FINAL REPORT

LGP Discrimination and Residual Risk Analysis on Standardized Test Sites—Camp Sibert and Camp San Luis Obispo

ESTCP Project MR-0811

JUNE 2010

Frank D. Francone
RML Technologies, Inc.

Dean A. Keiswetter
SAIC

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
LGP Discrimination and Residual Risk Analysis on Standardized Test Sites-Camp Sibert and Camp San Luis Obispo	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **RML Technologies, Inc,7606 S. Newland St,Littleton,CO,80128**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**130**	

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

-2LL	Minus two times Log Likelihood
11X	Depth corresponding to 11 times an object's diameter
AUC	Area under the Curve of a ROC curve
BRAC	Base Realignment and Closing
CFS	Correlation-Based Feature Selection
CNG	California National Guard
CWS	Chemical Warfare Service
DAQ	Data Acquisition Computer
DGM	Digital Geophysical Mapping
DoD	Department of Defense
EE/CA	Engineering Evaluation/Cost Analysis
EM61MTADS	EM61 MKII MTADS Array
EMI	Electromagnetic Induction
ESTCP	Environmental Security Technology Certification Program
FPU	Floating Point Unit
FUDS	Formerly Used Defense Site
GEMTADS	GEM-3 MTADS Array
GLRT	Generalized Likelihood Ratio Test
GPO	Geophysical Prove-Out
GPS	Global Positioning System
GSA	General Services Administration
HRR	Historical Records Review
IDA	Institute for Defense Analyses
IMU	Inertial Measurement Unit
LGP	Linear Genetic Programming
MAGMTADS	Magnetometer MTADS Array
MEC	Munitions and Explosives of Concern
MSEMS	Man Portable Simultaneous EMI and Magnetometer System
MRMR	Maximum Relevance Minimum Redundancy
MRS	Munitions Response Site
MTADS	Multi-sensor Towed Array Detection System

Nfa	Number of False Alarms
NOSLN	No On Site Learning Necessary
NRL	Naval Research Laboratory
Pclass	Probability of Correct Classification
PDF	Probability Density Function
RTK	Real Time Kinematic
QA	Quality Assurance
QC	Quality Control
RML	RML Technologies, Inc.
ROC	Receiver Operating Characteristic
SAIC	Science Applications International Corporation
SCORR	Slope Corrected
SERDP	Strategic Environmental Research and Development Program
SI	Site Investigation
SLO	San Luis Obispo
SNR	Signal to Noise Ratio
TEMTADS	Time Domain EM Discrimination Array
TOI	Targets of Interest
UTC	Universal Coordinated Time
UXO	Unexploded Ordnance

EXECUTIVE SUMMARY

This report describes a two-year UXO discrimination project at two sites: former Camp Sibert, Alabama and Camp San Luis Obispo ("SLO"), California. The demonstrations described in this report were performed under project Environmental Security Technology Certification Program (ESTCP) MM-0811 "Advanced MEC Discrimination Comparative Study on Standardized Test-Site Data Using Linear Genetic Programming (LGP) Discrimination." It was performed under the umbrella of the ESTCP Discrimination Study Pilot Program. The MM-0811 project demonstrates the application of the LGP Discrimination Process™ to the problem of UXO discrimination.

At the Camp Sibert site the objective was to discriminate potentially hazardous 4.2" mortars from non-hazardous shrapnel, range and cultural debris. Digital Geophysical Mapping ("DGM") was acquired by the ESTCP Program Office from a variety of sensor arrays.

At the SLO site, the objective was to discriminate a variety of potentially hazardous munitions, including 60mm mortars, 81mm mortars, 2.36 inch rockets, and 4.2" mortars from items that may be safely left in the ground.

The LGP Discrimination Process™ begins with the DGM from a site suspected of containing UXO. It then extracts attributes from potential targets in the DGM that may be UXO, uses Linear Genetic Programming ("LGP") and the attributes to rank the potential targets in their order of likelihood of being UXO, and finally, applies statistical residual risk analysis to determine which of the ranked targets may be safely left in the ground as Not-UXO.

The attributes extracted were analyzed by information-theoretic and statistical methods to reduce the attribute set to a small number of highly-predictive attributes. Then, Linear Genetic Programming (LGP) was used to rank the "blind" targets as either UXO or Not-UXO using a small "training" set of targets for which ground-truth was provided. Finally, statistical residual risk analysis was applied to the rankings and to the training ground-truth to determine the stop-digging cut-off.

At Camp Sibert predictions on a much larger "blind" data set containing one-hundred and nineteen seeded 4.2" mortars provided the metric for success. One-hundred percent of the UXO were located by the LGP discriminator with only a small number of false-positives. Near-perfect ROC curves (0.99+) were generated by the LGP-generated rankings. A high percentage of non-UXO were safely left in the ground as high-probability Not-UXO. A complete report on the results of the Camp Sibert study may be found in the ESTCP report date June 6, 2010 entitled *Interim Report—LGP Discrimination and Residual Risk Analysis on Standardized Test Sites—Camp Sibert (MM-0811*) (the "MM-0811 Interim Report").

At SLO, predictions on a larger "blind" data set comprised of 205 UXO and 1077 Not-UXO provided the metric for success. The LGP process generated a final ROC curve for items it ranked of 0.936 (1.0 is perfect) and all but three of the UXO appeared above the stop-digging threshold. At the stop-digging threshold, 35.9% of Not-UXO were left in the ground as high-probability Not-UXO. The SLO results are reported in the body of this report.

Finally, the intention in this project was to test an iterative process in which the first LGP rankings and risk analysis would be used to select further ground-truth. That further ground-truth would be used as the basis for additional LGP ranking and risk analysis. That process would

have iterated until a stop-digging decision was reached. The goal of iteration was to improve the ROC charts and to improve the accuracy of the stop-digging cutoff with additional ground-truth.

We were unable to demonstrate this iterative process at Camp Sibert because the ROC charts on our first iteration were near-perfect and the stop-digging thresholds accurately identified all UXO. Thus, there was little or no room for improvement with further iteration.

At SLO, the additional ground-truth on the second iteration significantly improved the discrimination over the first iteration discrimination by almost any metric. In other words, intelligently selecting which targets to “dig” and then rebuilding discrimination models using those new targets as training targets significantly improves discrimination performance.

1 INTRODUCTION

1.1 Background

In 2003, the Defense Science Board observed: "The … problem is that instruments that can detect the buried UXO also detect numerous scrap metal objects and other artifacts, which leads to an enormous amount of expensive digging. Typically 100 holes may be dug before a real UXO is unearthed! The Task Force assessment is that much of this wasteful digging can be eliminated by the use of more advanced technology instruments that exploit modern digital processing and advanced multi-mode sensors to achieve an improved level of discrimination of scrap from UXO . . . ." [1]

Significant progress has been made in classification technology over the past several years. To date however, testing of these approaches has been primarily limited to test sites with only limited application at live sites. Acceptance of these classification technologies requires demonstration of system capabilities at real UXO sites under real world conditions. Any attempt to declare detected anomalies to be harmless and requiring no further investigation will require demonstration to regulators of not only individual technologies, but an entire decision making process.

The FY06 Defense Appropriation contained funding for the "Development of Advanced, Sophisticated, Discrimination Technologies for UXO Cleanup" in the Environmental Security Technology Certification Program (ESTCP). ESTCP responded by conducting a UXO Classification Study at the former Camp Sibert, AL.[2] The results of this first demonstration were very encouraging. Although conditions were favorable at this site, a single target-of-interest (4.2-in mortar) and benign topography and geology, all of the classification approaches demonstrated were able to correctly identify a sizable fraction of the anomalies as arising from non-hazardous items that could be safely left in the ground. Of particular note, the contractor EM61-MK2 cart survey with analysis using commercially-available methods correctly identified more than half the targets as non-hazardous.

To build upon the success of the first phase of this study, ESTCP sponsored a second study in 2008 at a site with more challenging topography and a wider mix of targets-of-interest. A range at the former Camp San Luis Obispo, CA has been identified for this demonstration. This document describes the planned demonstration at San Luis Obispo.

1.2 Objective of the Demonstration

1.2.1 ESTCP UXO Discrimination Study

As outlined in the ESTCP UXO Discrimination Study Demonstration Plan[3], there are two primary objectives of this study:

---

[1] Department of Defense. (2003) *Report of the Defense Science Board Task Force on Unexploded Ordnance*.

[2] Nelson, H., Kaye, K., and Andrews, A. *ESTCP Pilot Program, Classification Approaches in Munitions Response,*

[3] ESTCP Program Office (2003) *Demonstration Plan for 2008 Classification Study V3*.

1. Test and validate detection and classification capabilities of currently available and emerging technologies on real sites under operational conditions.
2. Investigate in cooperation with regulators and program managers how classification technologies can be implemented in cleanup operations.

Within each of these two overarching objectives, there are several sub-objectives.

1. Test and evaluate capabilities by demonstrating and evaluating individual sensor and classification technologies and processes that combine these technologies. Compare advanced methods to existing practices and validate the pilot technologies for the following:
2. Detection of UXO
3. Identification of features that distinguish scrap and other clutter from UXO
4. Reduction of false alarms (items that could be safely left in the ground that are incorrectly classified as UXO) while maintaining Pds acceptable to all
5. Ability to identify sources of uncertainty in the classification process and to quantify their impact to support decision-making, including issues such as impact of data quality due to how data is collected
6. Quantify the overall impact on risk arising from the ability to clear more land more quickly for the same investment.
7. Include the issues of a dig-no dig decision process and related QA/QC issues
8. Understand the applicability and limitations of the pilot technologies in the context of project objectives, site characteristics, suspected ordnance contamination
9. Collect high-quality, well documented data to support the next generation of signal processing research

1.2.2 Specific Objectives of the Demonstration

Our objective is to advance and improve MEC discrimination performance by validating a decision process that (i) combines statistical analyses of digital geophysical mapping products and Linear Genetic Programming (LGP) methods to enable classification, and (ii) provides iterative quantitative residual risk assessments that may be used during the excavation phase to determine a stop-digging cutoff.

The objective was implemented in two phases—each associated with one year of the project.

1. **Camp Sibert Study**. The first year of this program (MM-0811) entailed UXO discrimination at former Camp Sibert in three tracks. The results of that study are reported in the MM-0811 Interim Report, on file with ESTCP.
2. **Camp San Luis Obispo (SLO) Study**. The second year of this program (MM-0811) entailed UXO discrimination and risk analysis at SLO. The results of that study are reported in the body of this report.

1.3 Regulatory Drivers

ESTCP has assembled an Advisory Group to address the regulatory, programmatic and stakeholder acceptance issues associated with the implementation of classification in the MR

process. Additional details can be found in the ESTCP UXO Discrimination Study Demonstration Plan.[4]

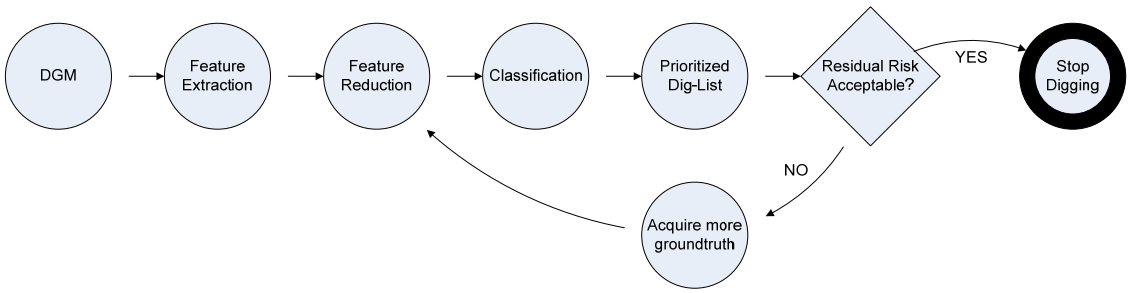
2 TECHNOLOGY

The LGP Discrimination Process is a multi-step, iterative process that uses Linear-Genetic-Programming (LGP) to perform the most difficult classification tasks and RML's Residual Risk Analysis is used to recommend a stop-digging decision for a customer-defined threshold. By describing it as an "iterative" process, we mean that early classification models are used to select ground-truth on which later models are trained.

When the iterative Residual Risk Analysis is added, the LGP Discrimination Process starts with a small training set for initial prioritization of targets. If the residual risk is too high to recommend a stop-digging decision, additional ground-truth is acquired and that ground-truth is added to the training set. From that larger training set, a better prioritized dig-list is built. This process continues until reaching a customer designated risk level for the probability that no intact MEC remain on the site.

Figure 1 shows the complete iterative process by which improved classification models are built as the site is excavated. The goal in the iteration is to characterize the tail of the probability density function that a given target is MEC as a function of dig-list ranking with the fewest possible number of excavations. From that tail, the residual risk of MEC remaining on site may be computed to customer specified confidence levels.

Figure 1. The LGP Discrimination Process including iterative residual risk analysis[5]



In Figure 1, the term "feature" is used in the same manner as the term "attribute" is used in the remainder of this report.

2.1 TECHNOLOGY DESCRIPTION

The steps in the LGP Discrimination Process are:

- Data Acquisition
- Data QAQC
- Attribute Extraction
- Attribute Reduction

[4] *Id.*

[5] We use the term "feature" in this figure to describe what is elsewhere in this report called an "attribute."

- Modeling
- Residual Risk Analysis
- Iterate. Request further Ground-truth and Iterate thru steps 4-6 until stop-digging decision is reached.

We will address each of these steps in this section.

2.1.1 Data Acquisition

The sensor used to collect data for the SLO project was MTADS EM61 MKII Array ("EM61MTADS"). The EM61MTADS was configured with four decay channels. In this report, we will refer to the first decay channel as "channel 1", the second decay channel as "channel 2" and so forth.

The Digital Geophysical Mapping ("DGM") generated by the EM61MTADS was used to perform discrimination and risk analysis on this project. This DGM data were provided to us by the ESTCP program office, leveled and lag corrected. So while data acquisition and preliminary leveling and lag correction is formally a step in our process, we did not perform data acquisition, leveling or lag correction as part of this project.

2.1.2 Data QAQC

The purpose of this step is to assure that the data on which we are performing modeling is good enough to support a no-dig decision for each target. Data QAQC is not a singular step that ends early in the process. It is an ongoing procedure that occurs throughout the LGP Discrimination Process.

Thus, we may determine that the DGM in the region of a target is sufficiently ambiguous or overlapping with an adjacent target that it may not be properly modeled. This would occur toward the beginning of our process and the principal tool used is visual examination of the gridded DGM. On the other hand, later, and after we have completed the Attribute Reduction step (see below), we observe the resulting distribution of attributes that have been identified as potentially important attributes. Statistical outliers on these attributes would be excluded from further analysis. Finally, after we perform residual risk analysis, we examine attribute space and may determine that the data density in attribute space is not sufficient to support a no-dig decision for a particular target.

The result of this process is that a certain portion of targets are assigned to a "cannot-analyze" category, which means that these targets must be dug regardless as they cannot be confidently designated as high-confidence Not-UXO.

2.1.3 Attribute Extraction

The purpose of attribute extraction is to measure numeric aspects of each target in a way that is meaningful to the ranking of UXO vs. Not-UXO. A reduced portion of those attributes become inputs to the LGP modeling algorithm. The attributes are statistics from the DGM, channel by channel, and statistics of ratios of adjacent channels in the region immediately surrounding the target. Those statistics are defined by ellipses and/or circles centered on the target—the ellipses being larger or smaller depending on the size of the target in the DGM. Thus, an ellipse was

defined for each target. The ellipse separates the region that comprises signal from the region that comprises background noise.

We extracted:

1. The first through third moments in each region for channels 1-4.
2. The first through third moments of the ratios of adjacent channels in each region.

2.1.4 Attribute Reduction

The Attribute Extraction process described above produces hundreds of statistics for every target. The goal in attribute reduction is to reduce the number of attributes used in modeling to just a handful of highly relevant attributes that contain complementary information content about the modeling problem.

We used a collection of tools at different points in the modeling process to reduce attributes. The purpose of this section is to introduce the tools generally. We will describe how they were applied to particular portions of this project as we address them individually. Typical application uses a variety of these techniques to reduce the number of attributes to a small number of candidates. The final selection is made using operator judgment, frequently based on analysis of graphic representations of attribute space and using quick and dirty modeling techniques to anticipate how a particular attribute set will perform on the ranking task. The particular sequence of application of the techniques is not pre-determined nor is every technique applied in every project. Rather, it is determined by the operator based on what sequence of techniques appear to be working well.

The goal is to generate a good reduced attribute set that will provide robust predictors. The goal is not to generate the optimal set, a problem that is computationally intractable. The techniques include:

2.1.4.1 Numeric Input Binning

Binning numeric variables is a fundamental technique in machine learning. We use two sorts of binning in this project. Binning is the process of assigning numeric values to discrete categories:

Equal-frequency Binning. In equal-frequency binning, a number of bins is specified and the numeric values are divided into that number of bins. This technique attempts to assign the same number of numeric values to each bin. Sometimes that is not entirely possible because of tied numeric values.

Chi-square Binning. Chi-square binning splits the numeric values into bins based on how well the splits do in minimizing the probability of Chi-square statistic of the 2x2 contingency table formed by the split of UXO and Not-UXO on either side of the bin boundary. This is a recursive technique. It starts by finding the single split that has the lowest probability. If the probability is greater than a selected parameter, binning stops. If it is less, then each bin is split in the same manner. Splitting continues in each bin partition until the probability is greater than the set probability parameter.

2.1.4.2 Mutual Information

Mutual Information between an independent variable and the dependent variable (UXO) is usually one of the first measures we look at. The mutual information of two discrete random variables *X* and *Y* may be computed as follows:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right)$$

We will refer to mutual information between a variable X and UXO as $I(UXO;X)$.

Typically, $I(UXO;X)$ is computed on a variable by variable basis and the results ranked. This gives a ranking of the variables that provide, by themselves, the most mutual information about the UXO/Not-UXO classification. This metric does not evaluate sets of attributes as predictors.

We compute *I* using discrete attributes and output. Accordingly, before any computation of *I*, it is necessary to bin the attributes first.

2.1.4.3 Symmetric Uncertainty

Symmetric uncertainty is a very close cousin of Mutual Information. Essentially it is Mutual Information adjusted for the bin count used to bin the variables. See Section 6.9.3.1.

2.1.4.4 Maximum Relevance Minimum Redundancy

Maximum Relevance Minimum Redundancy methods ("MRMR") locate attribute sets with the maximum amount of mutual information between the attribute set and the target output and simultaneously, the minimum amount of overlapping mutual information as between the individual attribute in the dataset.[6] In other words, MRMR does not look for just the best attributes measured by mutual information between the individual attributes and the target output. Such attributes are frequently highly correlated and contain very similar information about the target output. Having five such attributes adds little or nothing to our ability to solve the problem. Rather, MRMR attempts to construct the attribute set that collectively contains the most information about the target output. This is a first order computation—that is, it will not detect attributes that are important because of attribute interactions.

The MRMR algorithm is a greedy best-first algorithm. That is, it searches the entire attribute set for the single attribute that best increases the Relevance/Redundancy objective function. That attribute is added to the attribute set and that decision is not reexamined. Then the MRMR algorithm searches for the next attribute that, when added to the existing selected attribute set best maximizes the objective function. The size of the data set is passed to MRMR as a parameter and the algorithm returns the *n* best attributes using the MRMR criterion.

We compute MRMR attribute sets using discrete attributes and output. Accordingly, before any computation of MRMR, it is necessary to bin the attributes first.

---

[6] Hanchuan Peng, Fuhui Long, and Chris Ding (2005). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp.1226-1238.

2.1.4.5 Correlation Based Feature Selection

Correlation-Based Feature Selection ("CFS") is very similar to MRMR. Its goal is to derive attribute sets that, collectively, do a good job of predicting the target output.[7] The difference between CFS and MRMR is that CFS uses correlation coefficients instead of *I* as the measure of the predictive power of the attribute set and of the overlapping information included amongst the selected attributes. The advantage of CFS over MRMR is that it is not necessary to bin the attributes. The disadvantage is that CFS is not as good as MRMR at detecting non-linear relationships between attributes and the target output (UXO) and as between attributes selected for an attribute set.

We use CFS with a semi-greedy search algorithm. The algorithm adds the attribute that causes the largest gain in its objective function. However, unlike a purely greedy algorithm, our CFS algorithm is permitted to backtrack, that is, eliminate up to *n* of the most recently added attributes and start climbing from that spot. Obviously, if *n* is equal to the number of candidate attributes, then this is an exhaustive search algorithm, attempting all combinations of attributes.

2.1.4.6 Decision Trees

We use two forms of decision trees in variable reduction.

The first is the J48 single decision tree algorithm. It is an extension of the classic C4.5 decision tree algorithm.[8] J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. The J48 algorithm may be summarized as follows:

"J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurses on the smaller subsets created by the split. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling it to choose that class for all items in that node. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class."[9]

We use J48 as an alternative way to pick out attribute sets from MRMR and CFS. J48 is stronger at picking out interactions amongst attributes than is either MRMR or CFS.

Random Forests™ is a trademark of Leo Breiman. Random Forests is an ensemble decision tree algorithm that is reasonably fast and is does a good job of building preliminary models. We use Random Forests to assess the probable predictive result of a particular attribute set and also use its variable importance rankings as an attribute excluder. Random Forests is not particularly effective as an attribute includer.

---

[7] Hall M.A. (1998) "Correlation-based Feature Selection for Machine Learning." *Ph.D dissertation. Dept. of Computer Science, Waikato University,* 1998.

[8] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[9] http://www.opentox.org/dev/documentation/components/j48

2.1.4.7 Discipulus™ Input Impacts

After a project is finished, our core Discipulus Linear Genetic Programming software produces an "Input Impacts" report for that project. That report describes, for each attribute (input), what percentage of the best scoring evolved programs contained that attribute. It also measures how much each attribute contributes on average to the fitness of each of the thirty best evolved programs. We use these measures as attribute excluders.

2.1.5 Modeling

Modeling is the process of mapping the subset of attributes created in the attribute selection process to the ground-truth of UXO vs. Not-UXO. Our principal modeling tool is RML's Linear Genetic Programming ("LGP") software, Discipulus™ modified to use area under the curve of the ranking generated by an evolved program as the fitness function.

RML's LGP is an inductive-learning technology that is a variant of canonical Genetic Programming. Learning is conducted on a training dataset, consisting of an n-tuple for each Target, comprised of $n-1$ features that describe the Target and a class-label for the Target. The class-label, for MEC discrimination is, of course, whether the target is or is not MEC.

$r[1] = r[1] + x$
$r[1] = r[1] - 1$
$r[0] = r[1] * r[1]$
$r[1] = r[0] * r[1]$
$Output = r[0] + r[1]$

During training, LGP creates computer functions comprised of very simple Intel Floating Point Unit ("FPU"), machine-code instructions such as +,*,-,/, sqrt, power. Internal computations in the function operate directly on the FPU registers and the $n-1$ input features stored in memory. The LGP-created functions map the $n-1$ features to an output that orders the targets in terms of the likelihood they are MEC.

That ordering results in a prioritized dig-list. A simple five-line LGP function might look like the pseudo-code in the text box. (All registers are represented by $r[n]$ and are initialized to zero. The one input feature in this example is represented by x). This program uses two registers to represent a functional mapping of x to an output, $f(x)$. The function, in this case, is the polynomial, $f(x) = (x-1)^2 + (x-1)^3$.

LGP's learning algorithm has been described in detail in the literature.[10] In brief, LGP is a steady-state, evolutionary algorithm using tournament-selection to continuously improve a population of Intel machine-code functions. A single run is comprised of tournaments that compare the "fitness" of two randomly-selected programs that are repeated until a termination criterion is reached. At that point, the Intel machine-code of the selected best functions is decompiled into a readable and understandable C-code function. In practice, LGP is configured to perform many runs sequentially and to optimize its own parameters as those runs proceed.

For smaller training sets, we add noise to the inputs. The amount of noise is defined by a percentage parameter $noise\_\%$. The larger the $noise\_\%$ parameter, the wider the standard deviation of the added noise. The number of training instances is multiplied by another parameter, each instance having noise added to the inputs.

---

[10] Banzhaf, W., Nordin, P. Keller, R. Francone, F. (1998) *Genetic Programming, an Introduction*, Morgan Kaufman Publishers, Inc., San Francisco, CA at pp 257-264; and Nordin, J.P., Francone, F., and Banzhaf, W. (1999) "Efficient Evolution of Machine Code for CISC Architectures Using Blocks and Homologous Crossover," in *Advances in Genetic Programming 3*. Chapter 12 (MIT Press, Cambridge MA).

The *noise_%* parameter is set using k-fold cross-validation.[11]

The LGP models are then trained on data prepared using a technique called bagging, with the *noise_%* set to the previously selected value. Assume a training data set of size *n*. Bagging creates *j* separate training sets. Each training set is prepared by sampling rows from the training set *n* times with resampling.[12] A separate LGP model is trained from each bagged sample and the models are then applied to the blind data. The prediction for a blind target is the average ranking of each blind target by the multiple LGP models. On our dig list, the targets are ranked by those LGP predictions.

2.1.6 Residual Risk Analysis

Residual Risk Analysis is the final step of each iteration of the LGP Discrimination process. The goal of Residual Risk Analysis is to recommend a stop-digging decision based on the actual empirical results of applying the LGP Discrimination Process to a particular site, given a customer specified confidence level. The iterative process comprising the Residual Risk Analysis process is shown in Figure 1 and described generally in the text accompanying that figure.

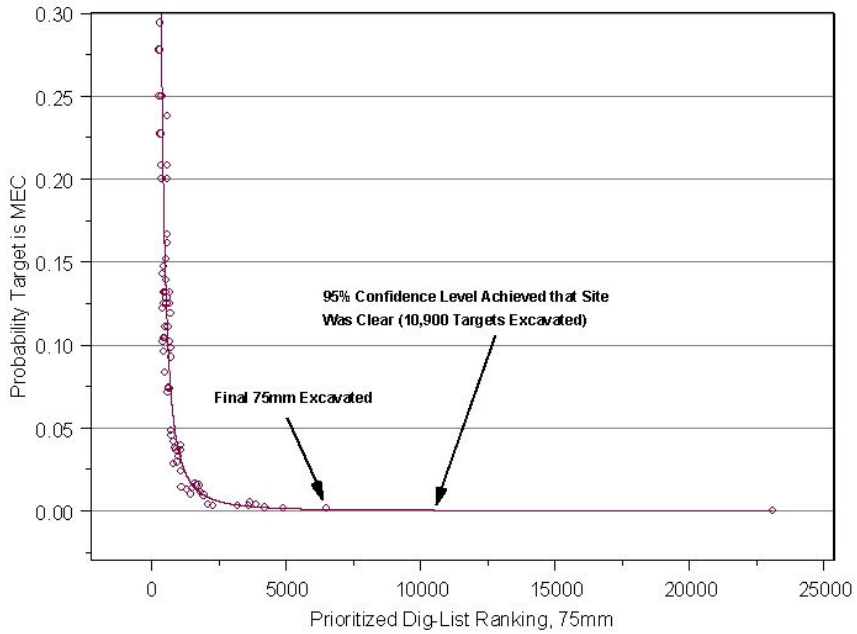
A key property of a prioritized dig-list that accurately discriminates UXO from other items is that the UXO are ranked nearer the start of the dig-list than clutter, hot-rocks, etc. As a result, *as excavation proceeds, the probability that the next item is MEC falls, not always continuously, but it falls.* Figure 2 shows that relationship in our work at F.E.Warren AFB. This is an example of a probability that falls relatively continuously as the dig-list ranking increases.

---

[11] Kohavi, Ron (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2 (12): 1137–1143. http://citeseer.ist.psu.edu/kohavi95study.html. (Morgan Kaufmann, San Mateo)

12 Breiman, L. (1996). “Bagging Predictors”. *Machine Learning* 24 (2): 123–140.

Figure 2. Relationship between prioritized dig-list ranking and probability that a target was 75mm UXO at F.E.Warren AFB.



The circles represent a measured probability that Targets were MEC in the vicinity of each MEC item found. The falling probability as Rank increases is clearly shown. The red line in Figure 2 is the maximum likelihood power-law relationship fit to these data after ranking 278. (The fit started there because the linear portion of the log-log transformed data in these data started at ranking 278.)

A classifier that produces a high-quality ROC chart will always have the property of falling empirical probability as rank increases on the dig-list. In this residual risk step, we fit an appropriate, simple model to the declining probability of UXO as a function of dig-list rank. Rank is calculated using the predictive scores output by LGP and the scores are combined across training and blind data to create a common ranking metric for the two data sets. Candidates for the most appropriate model that we considered in this project are Power Law fit, Exponential fit, Kernel Regression fit or Logistic Regression fit.

Once the model is fit on labeled, training data, we predict the probability of UXO as a function of rank for unlabeled, blind data. From those probabilities, we also predict the residual risk of UXO. That residual risk is the probability that any sequence of targets from the *nth* ranked target on the dig-list to the maximum ranked target on the dig list contain one or more UXO. For the *nth* ranked target, we compute that residual risk probability as the OR of the probabilities of UXO for all targets from the *nth* ranked target to the maximum ranked target. Thus, at any given target ranking, the risk remaining (probability) that the targets with a *higher* ranking (less likely to be UXO) contain one or more UXO items is the OR of the probabilities for all higher ranking targets.

The OR operator when applied to the probabilities of two events labeled A and B (for example, target A OR target B being UXO), is computed as follows:

Equation 1:

$$P(A\_OR\_B) = P(A) + P(B) - P(A\_AND\_B)$$

In Equation 1, we use $P(A\_and\_B) = P(A) \cdot P(B)$.[13]

In the present study, the above formula is applied to all targets ranked to the right of the plotted rank (that is, ranked less likely to be UXO) by chaining the computation. This is applied as follows: Assume that three targets have a higher ranking than a given rank and that the targets are labeled A, B, and C. Given the definition of *P(A_OR_B)* in Equation 1 above, we can now compute the probability of A OR B OR C as follows:

Equation 2:

$$P(A\_OR\_B\_OR\_C) = P(A\_OR\_B) + P(C) - P((A\_OR\_B)\_AND\_C)$$

Equation 2 may be expanded to compute the OR value for the probability that at least one of any number of targets is UXO.[14] When we compute the residual risk for rank *n*, we expand this equation to include all probabilities for all targets ranked greater than *n*.

Thus, at each rank on the dig-list, we measure the residual risk using this OR of probabilities computation. The key point here is that the probabilities used in our Residual Risk Analysis are based on the actual, site-specific empirical results of applying the LGP-based dig-list to the site.

2.1.7 Iteration

At each risk analysis step, and based on the ground-truth at that time, we estimate the Target parameters described above using the LGP discrimination process (resulting in a prioritized dig-list) and determine if a stop-digging decision is warranted at the specified confidence level. If not, we request more ground-truth, re-estimate the parameters using all ground-truth then available, and determine (based on the new estimates) if a stop-digging decision is warranted. That process continues until a stop-digging decision is warranted at the specified confidence level.

2.2 TECHNOLOGY DEVELOPMENT

This technology has not been previously developed under grant from ESTCP.

2.3 ADVANTAGES AND LIMITATIONS OF THE TECHNOLOGY

Key differences between LGP and other learning algorithms are:

LGP does not just derive parameters for a specified functional form—it derives the functional form itself and optimizes the parameters of the derived functional form, in one pass;

Because LGP software operates directly on populations comprised of Intel machine code functions, it is approximately two orders of magnitude faster than comparable inductive-learning

[13] Kachigan, S. (1986) *Statistical Analysis*, Radius Press, NY, NY.

[14] This computation assumes that the probability of UXO for target *j* and the probability of UXO for target *j+1* are independent.

technologies.[15] Coupled with the fact that this software can run on multiple CPU's over a network in parallel, LGP is capable evaluating millions of functions on large data sets in commercially reasonable time-frames;

LGP software has been subjected to extensive in-house and third-party testing on a wide variety of data sets over a nine-year period. Results have been published by RML and SAIC[16] and by third-parties[17];

LGP was designed to prevent, insofar as possible, building models of the training-set noise rather than the signal sought to be modeled. (LGP's resistance to fitting noise has been noted in the literature); and

The version of Discipulus used in this project uses as its fitness function, the Area under the curve ("AUC") of the ROC curve defined by the evolved program ranking. In other words, the evolution process is geared toward creating a good ranking. Most other inductive learning algorithms perform some kind of classification and then convert that into a ranking. This is a subtle but important difference because classifying items as, say, UXO vs. Not-UXO is a different goal than ranking them well. Discipulus produces much better rankings when it uses an AUC fitness function than it does when using a classification fitness function.

The principal disadvantage of LGP is that it requires experienced data modelers for its operation. It is a very powerful modeling tool because of the breadth of the search it can conduct over a very large solution space—both because of its speed and because it evolves functional form, not just parameterization of a preexisting functional form. If used improperly, it can produce wonderful-looking results on known data and very poor results when applied to new data.

---

[15] Banzhaf, W., Nordin, P. Keller, R. Francone, F. (1998) *Genetic Programming, an Introduction*, Morgan Kaufman Publishers, Inc., San Francisco, CA at pp 257-264; and Nordin, J.P., Francone, F., and Banzhaf, W. (1999) "Efficient Evolution of Machine Code for CISC Architectures Using Blocks and Homologous Crossover," in *Advances in Genetic Programming 3*. Chapter 12 (MIT Press, Cambridge MA); and Fukunaga, A., Stechert, Mutz, D. (1998) "A Genome Compiler for High Performance Genetic Programming," in: *Proceedings of the Third Annual Genetic Programming Conference*, Jet Propulsion Laboratories, California Institute of Technology Pasadena, CA, Morgan Kaufman Publishers, pp. 86–94.

[16] Several years of comparative studies by RML and SAIC are reported in: Francone, F. D., and Deschaine, L.M., (2004) *Extending the Boundaries of Design Optimization by Integrating Fast Optimization Techniques with Machine-Code-Based Linear Genetic Programming, Information Sciences Journal—Informatics and Computer Science*, Elsevier Press, Vol. 161/3-4 pp 99-120 (see sections 8.3-8.6 for results of the comparative study) Amsterdam, the Netherlands. In brief summary, RML's LGP software consistently performs as well as the best-tested alternative classification algorithms or better, on blind data. Other learning algorithms sometimes perform as well as RML's LGP algorithm but are not nearly as consistent as RML's LGP in producing high-quality results on unseen, testing data.

[17] See: (1) Mukkamala, S., Sung, A., Abraham, A., (2004) "Modeling Intrusion Detection Systems Using Linear Genetic Programming Approach," in *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*; (2) S. Mukkamala, Q. Liu, R. Veeraghattam, A. H. Sung (2005) "Computational Intelligent Techniques for Tumor Classification (Using Micro array Gene Expression Data)." *International Journal of Lateral Computing*, Vol.2, No. 1, ISSN 0973-208X, pp. 38-45; and (3) Mukkamala, G. D. Tilve, A. H. Sung, B. Ribeiro, A . S. Vieira (2006) Computational Intelligent Techniques for Financial Distress Detection. *International Journal of Computational Intelligence Research.*

3 PERFORMANCE OBJECTIVES

The relevant objectives include (i) Target-of-Interest retention rate, (ii) non-Target-of-Interest reduction rate, (iii) ; (iv) Minimize number of target that cannot be analyzed; and (v) Minimize the number of blind targets sampled. The focus will be on identifying items that may be safely left in the ground. The main failure is misclassifying a target of interest as an item that can be left in the ground.

Items that may be safely left in the ground shall include HE fragments, single fins, cultural debris and geology.

Table 1. Performance objectives summary

Performance Objective	Metric	Data Required	Success Criteria	Result
Maximize correct classification of munitions	Number of targets-of-interest retained.	Prioritized anomaly lists Scoring reports from IDA	Approach correctly classifies 100% targets-of-interest	Correctly classified 98.6% of targets of interest
Maximize correct classification of non-munitions	Number of false alarms eliminated.	Prioritized anomaly lists Scoring reports from IDA	Reduction of false alarms by > 30% while retaining all targets of interest	False alarm rate reduced by 28.4% while retaining all targets of interest
Specification of no-dig threshold	P_{class} and N_{fa} at demonstrator operating point.	Demonstrator -specified threshold Scoring reports from IDA	Threshold specified by demonstrator to achieve criteria above	98.6% of targets-of-interest correctly classified. False alarm rate reduced by 35.9%.
Minimize number of anomalies that cannot be analyzed	Number of anomalies that must be classified as "Unable to Analyze."	Demonstrator target parameters	Reliable target parameters can be estimated for > 90% of anomalies.	Reliable target attributes were estimated for 82% of targets
Minimize the number of blind targets sampled	Number of targets initially classified as blind data that are sampled in the second and subsequent iterations	Requests for ground-truth on second and subsequent iterations Initial blind data list	Requested Ground-truth for sampling does not exceed 20% of initial blind targets in the aggregate	20% of blind targets sampled

The following sections provide a more detailed description of these objectives.

3.1 Objective: Maximize correct classification of munitions

This is one of the two primary measures of the effectiveness of this approach. By collecting high-quality data and analyzing those data with advanced feature extraction and classification

algorithms we expect to be able to classify the targets with high efficiency. This objective concerns with the component of the classification problem that involves correct classification of items-of-interest.

3.1.1 Metric

The metric for this objective is the number of items on the master anomaly list that can be correctly classified as munitions by each classification approach.

3.1.2 Data Requirements

A prioritized dig list for the targets on the master anomaly list will be prepared for each of the data sets analyzed as part of this demonstration. IDA personnel will use their scoring algorithms to assess the results.

3.1.3 Success Criteria

The objective will be considered to be met if all of the items-of-interest are correctly labeled as munitions on the prioritized anomaly list.

3.1.4 Result

At the demonstrated designated stop-digging threshold on our final dig list, 98.6% of munitions were classified as munitions.

3.2 Objective: Maximize correct classification of NON-munitions

This is the second of the two primary measures of the effectiveness of this approach. This objective relates to the component of the classification problem that involves false alarm reduction.

3.2.1 Metric

The metric for this objective is the number of items-of-interest on the master dig list that can be correctly classified as non-munitions by each classification approach.

3.2.2 Data Requirements

A prioritized dig list for the targets on the master anomaly list will be prepared for each of the data sets analyzed as part of this demonstration. IDA personnel will use their scoring algorithms to assess the results.

3.2.3 Success Criteria

The objective will be considered to be met if more than 30% of the non-munitions items can be correctly labeled as non-munitions while retaining all of the targets-of-interest on the dig list.

3.2.4 Result

At 100% Pd, 28.4% of the blind non-munitions were correctly classified as non-munitions.

3.3 Objective: Specification of no-dig threshold

In a retrospective analysis as will be performed in this demonstration, it is possible to tell the true classification capabilities of a classification procedure based solely on the prioritized dig list

submitted by each demonstrator. In a real-world scenario, all targets may not be dug so the success of the approach will depend on the ability of an analyst to accurately specify their dig/no-dig threshold.

3.3.1 Metric

P_{class} and number of false alarms, N_{fa}, at the demonstrator-specified threshold are the metrics for this objective.

3.3.2 Data Requirements

A ranked anomaly list with a dig/no-dig threshold indicated will be prepared for each of the data sets analyzed as part of this demonstration. IDA personnel will use their scoring algorithms to assess the results.

3.3.3 Success Criteria

The objective will be considered to be met if more than 30% of the non-munitions items can be correctly labeled as non-munitions while retaining all of the targets-of-interest at the demonstrator-specified threshold.

3.3.4 Result

At the stop-digging threshold, 35.9% of all blind non-munitions items were correctly classified as non-munition.

3.4 Objective: Minimize number of anomalies that cannot be analyzed

Anomalies for which reliable parameters cannot be estimated cannot be classified by the classifier as high confidence non-munitions. These anomalies must be placed in the dig category and reduce the effectiveness of the classification process.

3.4.1 Metric

The percentage of anomalies for which reliable parameters cannot be estimated is the metric for this objective.

3.4.2 Data Requirements

A list of all target parameters along with a list of those anomalies for which parameters could not be reliable estimated will be submitted for each of the data sets analyzed as part of this demonstration.

3.4.3 Success Criteria

The objective will be considered to be met if reliable parameters can be estimated for > 85% of the anomalies on each sensor anomaly list.

3.4.4 Result

82% of blind targets were assessed as having sufficiently good data for classification. The rest of the blind targets were assessed as cannot-analyze.

3.5 Objective: Minimize the Number of Blind Targets Sampled

This is an iterative process that involves demonstrator specified sampling of blind targets. The goal was to not exceed the specified percent of additional samples.

3.5.1 Metric

Percent of targets initially classified as blind data that are sampled in the second and subsequent iterations Data Requirements

3.5.2 Data Requirements

Demonstrator's requests for further ground-truth and original master dig-list.

3.5.3 Success Criteria

Requested ground-truth for sampling does not exceed 20% of initial blind targets in the aggregate.

3.5.4 Result

Demonstrator requested 20% of initial blind targets as samples.

4 SITE DESCRIPTION

The site description material reproduced is here is taken from the recent SI report[18]. More details can be obtained in the report. The former Camp San Luis Obispo is approximately 2,101 acres situated along Highway 1, approximately five miles northwest of San Luis Obispo, California. The majority of the area consists of mountains and canyons. The site for this demonstration is a mortar target on hilltop in Munitions Response Site (MRS) 05 (within former Rifle Range #12).

4.1 Site Selection

This site was chosen as the next in a progression of increasingly more complex sites for demonstration of the classification process. The first site in the series, Camp Sibert, had only one target-of-interest and item "size" was an effective discriminate. At this site, there are at least four targets-of-interest: 60-mm, 81-mm, 4.2-in mortars and 2.36-in rockets.

4.2 Site History

Camp San Luis Obispo was established in 1928 by California as a National Guard Camp. Identified at that time as Camp Merriam, it originally consisted of 5,800 acres. Additional lands were added in the early 1940s until the acreage totaled 14,959. During World War II, Camp San Luis Obispo was used by the U.S. Army from 1943 to 1946 for infantry division training including included artillery, small arms ranges, mortar, rocket, and grenade ranges. According to the Preliminary Historical Records Review (HRR), there were a total of 27 ranges and thirteen

[18] Parsons, Inc. (September 2007). *Final Site Inspection Report, Former Camp San Luis Obispo, San Luis Obispo, CA.*

training areas located on Camp San Luis Obispo during World War II. Construction at the camp included typical dwellings, garages, latrines, target houses, repair shops, and miscellaneous range structures. Following the end of World War II, a small portion of the former camp land was returned to its former private owners. The U.S. Army was making arrangements to relinquish the rest of Camp San Luis Obispo to the State of California and other government agencies when the conflict in Korea started in 1950. The camp was reactivated at that time.

The U.S. Army used the former camp during the Korean War from 1951 through 1953 where the Southwest Signal Center was established for the purpose of signal corps training. The HRR identified eighteen ranges and sixteen training areas present at Camp San Luis Obispo during the Korean War. A limited number of these ranges and training areas were used previously during World War II. Following the Korean War, the camp was maintained in inactive status until it was relinquished by the Army in the 1960s and 1970s. Approximately 4,685 acres was relinquished to the General Services Administration (GSA) in 1965. GSA then transferred the property to other agencies and individuals beginning in the late-1960s through the 1980s; most of which was transferred for educational purposes (Cal Poly and Cuesta College). A large portion of Camp San Luis Obispo (the original 5,880 acres) has been retained by the California National Guard (CNG) and is not part of the FUDS program

4.3 Site Topography and Geology

The Camp San Luis Obispo site consists mainly of mountains and canyons classified as grassland, wooded grassland, woodland, or brush. A major portion of the site is identified as grassland and is used primarily for grazing. Los Padres National Forest (woodland) is located to the north-northeastern portion of the site. During the hot and dry summer and fall months, the intermittent areas of brush occurring throughout the site become a critical fire hazard.

The underlying bedrock within the Camp San Luis Obispo site area is intensely folded, fractured, and faulted. The site is underlain by a mixture of metamorphic, igneous, and sedimentary rocks less than 200 million years old. Scattered throughout the site are areas of fluvial sediments overlaying metamorphosed material known as Franciscan mélange. These areas are intruded by plugs of volcanic material that comprise a chain of former volcanoes extending from the southwest portion of the site to the coast. Due to its proximity to the tectonic interaction of the North American and Pacific crustal plates, the area is seismically active.

A large portion of the site consists of hills and mountains with three categories of soils occurring within: alluvial plains and fans; terrace soils; and hill/mountain soils. Occurring mainly adjacent to stream channels are the soils associated with the alluvial plains and fans. Slope is nearly level to moderately sloping and the elevation ranges from 600 to 1,500 feet. The soils are very deep and poorly drained to somewhat excessively drained. Surface layers range from silty clay to loamy sand. The terrace soils are nearly level to very steep and the elevations ranges from 600 to 1,600 feet. Soils in this unit are considered shallow to very deep, well drained, and moderately well drained. The surface layer is coarse sandy loam to shaley loam. The hill/mountain soils are strongly sloping to very steep. The elevation ranges from 600 to 3,400 feet. The soils are shallow to deep and excessively drained to well-drained with a surface layer of loamy sand to silty clay.

4.4 Munitions Contamination

A large variety of munitions have been reported as used at the former Camp San Luis Obispo. Munitions debris from the following sources was observed in MRS 05 during the 2007 SI:

- 4.2-inch white phosphorus mortar
- 4.2-inch base plate
- 3.5-inch rocket
- 37mm
- 75mm
- 105mm
- 60mm mortar
- 81mm mortar
- practice bomb
- 30 cal casings and fuzes.
- flares found of newer metal; suspected from CNG activities

At the particular site of this demonstration, 60-mm mortars, 81-mm mortars, 2.36 inch rockets and 4.2-in mortars and mortar fragments have been observed. The excavation of two grids as part of the preparatory activities provided information on the munitions at this target site.

4.5 Site Geodetic Control Information

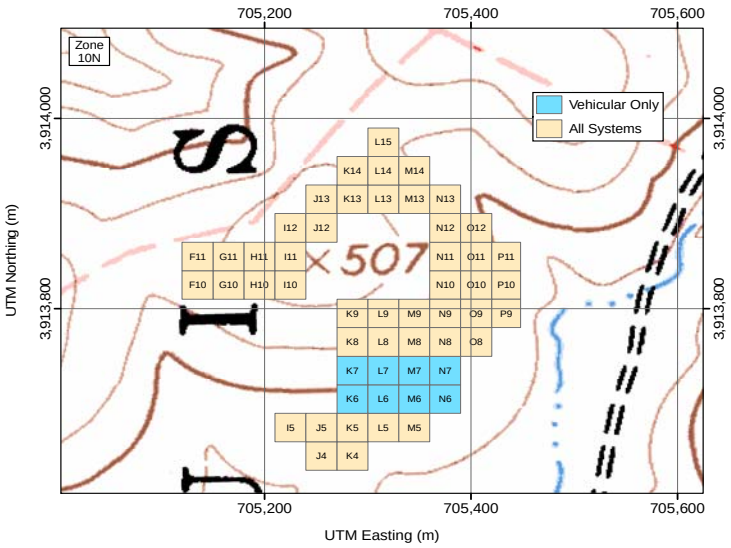
Figure 3. Geodetic Control at the former Camp San Luis Obispo site

ID	Latitude	Longitude	Elevation (m)	Northing (m)	Easting (m)	HAE (m)
ESTCP	35º 20' 37.77465" N	120º 44' 25.95073"W	113.69	3,913,515.94	705,330.89	76.01

4.6 Site Configuration

The demonstration site was configured as one 11.8-acre area. Details of the final site extent are shown in Figure 4. The calibration strip and training pit were located off the site, convenient to the access road.

Figure 4. Final layout of the demonstration site showing the grids to be surveyed by all systems (10 acres) and the additional 8 grids (1.8 acres) to be surveyed by the vehicular systems.



5 TEST DESIGN

5.1 CONCEPTUAL EXPERIMENTAL DESIGN

The overall objective of this ESTCP project was to demonstrate a methodology for the use of classification in the munitions response process. The three key components of this methodology are collection of high-quality geophysical data and principled selection of anomalous regions in those data, analysis of the selected anomalies to extract target attributes and the use of those attributes to construct a prioritized dig list.

The ESTCP Program Office coordinated data collection activities. This included all preparatory activities, arranging for a data collection by well-validated systems, selection of anomalies for analysis from each geophysical data set, and compilation of the individual sensor anomaly lists into a master list.

Validation digging was also coordinated by the Program Office. Because this is a demonstration, all anomalies on the master dig list were investigated. The underlying targets were uncovered, photographed, located with a cm-level GPS system, and removed. The identities of a small number of the recovered items plus the digital geophysical mapping (DGM) were provided to the demonstrator for use as training data. The identities of the remainder of the targets were retained by the program office as “blind” data to validate demonstrator’s results.

The demonstrator received and processed NRL’s EM61MTADS array data to extract attributes for each anomaly. The project was to proceed iteratively. Demonstrator would produce a prioritized dig-list, a stop-digging threshold and a probability that any UXO remained on the site, given the then known ground-truth and the stop-digging threshold. Demonstrator would then request further ground-truth, produce a new dig-list and stop-digging threshold, given the total known ground-truth. Demonstrator expected and performed two such iterations.

At the conclusion of each iteration of training, demonstrator would submit LGP-based prioritized dig list based on the EM61MTADS data. The list was ordered from the item that is most likely not hazardous (Not-UXO) through the item that is most likely munitions (UXO). The anomalies for which demonstrator was not able to extract meaningful parameters were placed at the bottom of the list.

These dig-lists were scored by the Institute for Defense Analyses with emphasis on the number of items that are correctly labeled non-hazardous while correctly labeling all munitions items.

The primary objective of the demonstration was to assess how well demonstrator was able to order the prioritized anomaly list and specify the threshold separating high confidence clutter from all other items. The secondary objective will be to determine the classification performance that could be achieved by each approach through a retrospective analysis.

5.2 SITE PREPARATION

Prior to the start of the surveys, the site will have been seeded with the items of interest under the guidance of the Program Office Seeding Plan. A Calibration Strip containing two of each item of interest and a selection of canonical objects (e.g. metal spheres) will be installed near the demonstration site and the site logistics location. One GPS control point is available on site. Basic facilities such as portable toilets, storage container, and generators for power are not available on its and will be mobilized in prior to the start of the survey.

5.3 DATA ACQUISITION SYSTEM SPECIFICATIONS

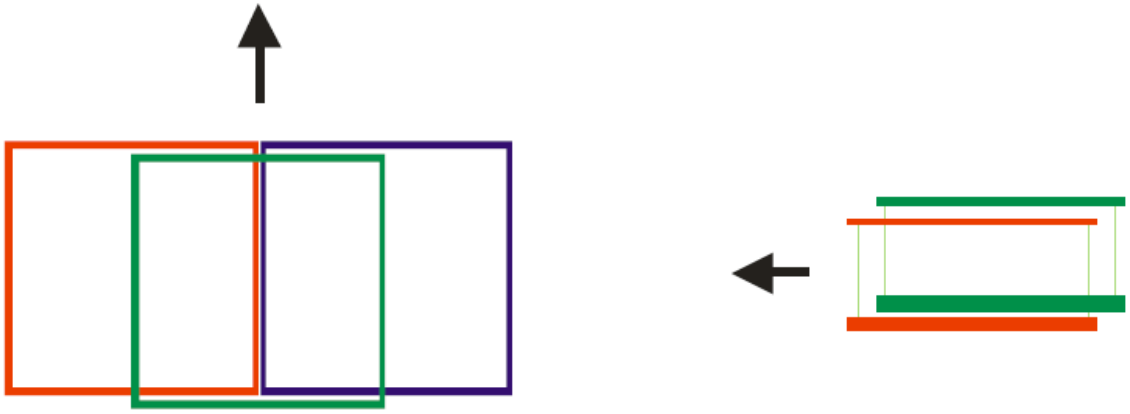
This demonstration was conducted with data from the EM61MTADS tow vehicle and subsystems. The sections below will address the EM61MTADS sensors and pilot guidance system.

5.3.1 EM61 MkII Array

The EM61MTADS array is an overlapping array of three pulsed-induction sensors specially modified by Geonics, Ltd. based on their EM61 MkII sensor with 1m x 1m sensor coils.

The array configuration is shown schematically in Figure 5. The direction of travel for the array is indicated by the black arrows. Sensors #1 (Red) and #3 (Blue) are mounted side by side on the trailer while Sensor #2 (Green) is mounted 8 cm above and 10 cm aft of the other two sensors. Each EM61 MkII sensor is composed of a bottom coil and a top coil separate by fiberglass standoffs. The nominal ride height of the bottom coils is 33.5 cm above the ground and the top coil is mounted 43.5 cm above the bottom coil (bottom of coil to bottom of coil separation). The bottom coil is 5.5 cm tall and the top coil is 2.5 cm tall.

Figure 5. Top and side schematic views of the EM61MTADS array.



The EM61 MkII sensors employed by MTADS have been modified to make them more compatible with vehicular speeds and to increase their sensitivity to small objects. The array is operated with the three transmitters synchronized to generate the largest transmit moment. The sensor repetition rate is 125 Hz, corresponding to a period of 8 ms. The transmit pulse is approximately 2.9 ms long, approximately $250\,A\times m^2$, and turns off in approximately $50\,\mu s$. The EM61 MkII sensor can be operated in one of two modes: 1) in 4-channel ("4") mode, in which 4 time gate measurements are made for the bottom coil or 2) in Differential ("D") mode, in which 3 time gate measurements are made for the bottom coil, and one is made for the top coil. The timing of the gates has been altered and the delay times are given in Table 2.

Table 2. NRL EM61 MkII Gate timing parameters

Channel	4 Gate Mode	Delay (μs)	Differential Mode	Delay (μs)
1	Bottom Coil	307	Bottom Coil	307
2	Bottom Coil	508	Top Coil	307
3	Bottom Coil	738	Bottom Coil	738
4	Bottom Coil	1000	Bottom Coil	1000

The notation "Channel 1" or "first decay channel" for time gate 1, "Channel 2" or "second decay channel" for time gate 2 and so forth is used in the remainder of this document.

EM61MTADS surveys have typically been performed using the Differential mode. As a consensus decision between the Program Office and all of the demonstrators involved in the UXO Classification Study, the 4-channel mode was used for this demonstration.

While the output data packet format is identical to that of the standard MkII instrument as given in the Geonics EM61 MkII manual, there are some important differences in the interpretation. First, as mentioned above, the time gate delay times have been altered. Second, the byte order for the time gate Scale Factors is gates 1,4,3,2 rather than the typical 1,2,3,4. The data channels are also presented in the order gates 1,2,3,4 for 4-gate mode, or gates 1,D,3,4 for differential mode. All conversions from raw counts to response in mV are given as:

$$RESPONSE = \frac{DATAx4.8333}{RANGE}$$

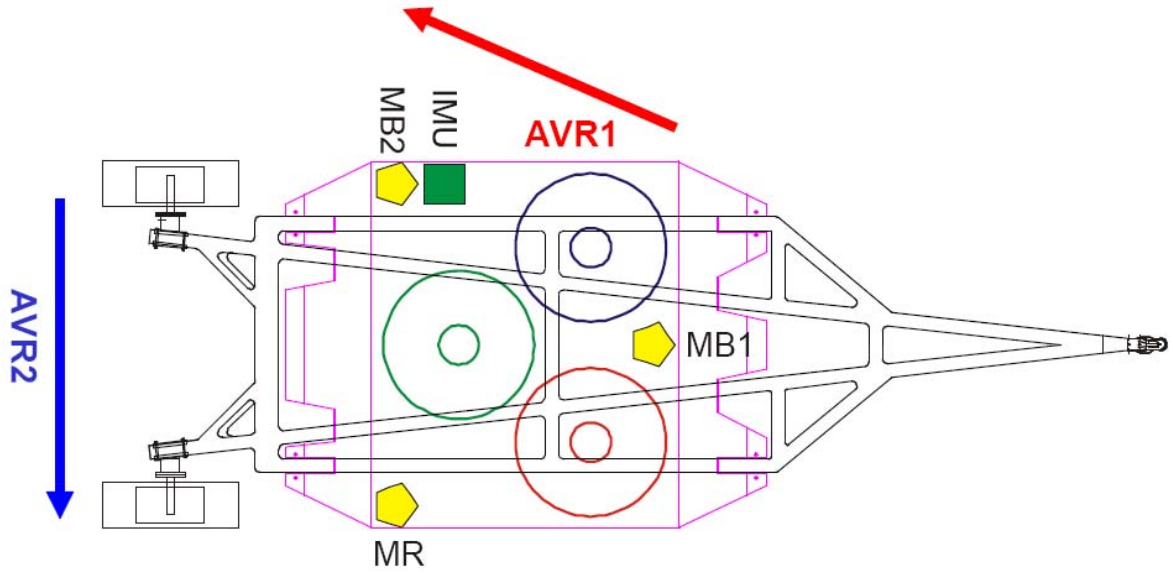
The channel-specific *RANGE* values are 100, 10, or 1, as indicated in the Scale Factor parameter in the raw data packet. Nominal survey speed is 3 mph and the sensor readings are recorded at 10 Hz. This results in a down-track sampling of ~15 cm and a cross-track interval of 50 cm. In order to obtain sufficient "looks" at the anomalies, or to insure illumination of all three principle axes of the anomaly with the primary field, data is collected in two orthogonal surveys. The EM61MTADS array being pulled by the MTADS tow vehicle is shown in Figure 6.

Individual sensors in the EM61MTADS array are located using a three-receiver RTK GPS system shown schematically in Figure 7. The three-receiver configuration extends the concept of RTK operations from that of a fixed base station and a moving rover to moving base stations and moving rovers. The lead GPS antenna (and receiver, MB1) receive corrections from the fixed base station at 1 Hz in the same manner as for the magnetometer MTADS. This corrected position is reported at 10-20 Hz using a vendor-specific National Marine Electronics Association (NMEA) NMEA-0183 message format (PTNL,GGK or GGK). The MB1 receiver also operates as a 'moving base,' transmitting corrections (by serial cable) to the next GPS receiver (MB2), which uses the corrections to operate in RTK mode.

Figure 6. EM61MTADS array pulled by the MTADS tow vehicle.

A vector (AVR1, heading (yaw), angle (pitch), and range) between the two antennae is reported at 10 Hz using a vendor-specific NMEA-0183 message format (PTNL,AVR or AVR). MB2 also provides 'moving base' corrections to the third GPS antenna (MR) and a second vector (AVR2) is reported at 10 Hz. All GPS measurements are recorded at full RTK precision, ~2-5 cm. All sensor readings are referenced to the GPS 1-PPS output to fully take advantage of the precision of the GPS measurements. An Inertial Measurement Unit (IMU) is also included on the sensor array to provide complimentary platform orientation information. The IMU is a Crossbow VG300 running at 30 Hz.

Figure 7. MTADS EM trailer with approximate locations of GPS and IMU equipment indicated.



The colored circles represent the GEM-3 sensors of the GEMTADS array (not involved in this project).

A close-up view of the sensor platform is shown in Figure 8 which shows the three GPS antennae and the IMU (black box under the aft port GPS antenna). The airborne adjunct of the MTADS, the AMTADS uses a similar configuration with two GPS antennae/receivers to provide the yaw and roll angles of the sensor boom and pitch from the IMU.

Figure 8. Close-up of EM61MTADS array with GPS and IMU.

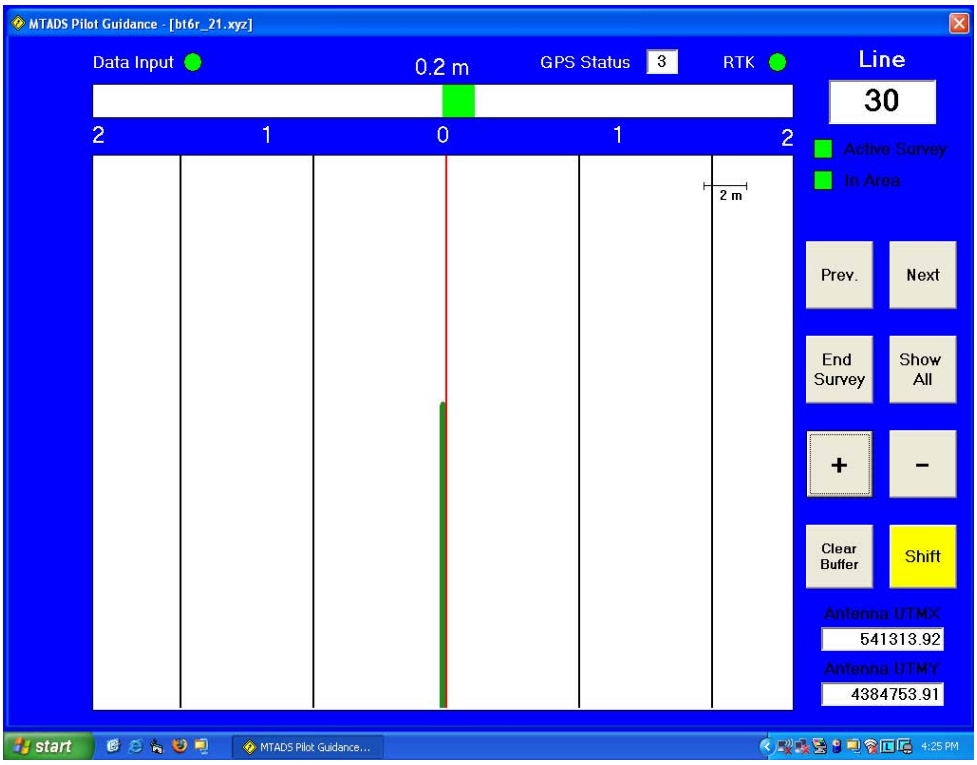
The individual data streams (sensor readings, GPS positions, times, etc.) are collected by the data acquisition computer, running the MagLogNT software package, and are each recorded in a separate file. These individual data files, which share a root name, consist of three EM61 MkII sensor data files, four GPS files (one containing the GGK and the first AVR sentences, another containing the second AVR sentence, a third containing the UTC time tag, and the fourth containing the computer time-stamped arrival of the GPS 1-PPS), and one IMU file. The EM61

MkII and IMU data files are recorded in packed binary formats. All GPS files are ASCII format. All these files are transferred to the data analyst using magnetic disks.

5.3.2 Pilot Guidance System

The GPS positioning information used for data collection is shared with an onboard navigation guidance display and provides real-time navigational information to the operator. The guidance display was originally developed for the airborne adjunct of the MTADS system (AMTADS) and is installed in the vehicle and available for the operator to use. Figure 9 shows a screenshot of the guidance display configured for vehicular use. An integral part of the guidance display is the ability to import a series of planned survey lines (or transects) and to guide the operator to follow these transects. In the context of this demonstration, the pilot guidance display can be used to guide the operator to the survey area and provide immediate feedback on progress and data coverage. The display provides a left right course correction indicator, an optional altitude indicator for aircraft applications, and color-coded flight swath overlays where the current transect is displayed in red and the other transects are displayed in black for operator reference. The survey course-over-ground (COG) is plotted for the operator in real time on the display. The COG plot is color-coded based on the RTK GPS system status. When fully operational, the COG plot is color-coded green. If the system status is degraded, the COG plot color changes from green to yellow to red (based on severity) to warn the operator and allow for on-the-fly reacquisition of the affected area. Figure 9 shows the operator surveying line 30 of a transect plan.

Figure 9. Screenshot of MTADS Pilot Guidance display.



5.4 CALIBRATION ACTIVITIES FOR SENSOR

5.4.1 Sensor Calibration

For the EM61MTADS array, the standard performance checks include three types of measurements. At the beginning of field work and again each morning quiet, static data are collected for a period (15 - 20 minutes or as directed by the Quality Assurance Office (QAO)) with all systems powered up and warmed up (typically 30 minutes after the transmitter is turned on). Next, a calibration item, a 4" diameter Aluminum (Al) sphere, is placed in well-defined positions along a fiberglass rail mounted a fixed distance above the array to verify the spatial response of the array to the object. The system is stationary for this data collection. Finally, a systems timing check using a fixed-position wire or chain placed on the ground is conducted. At the discretion of the QAO, the timing check may be repeated in the middle of the survey day. At the discretion of the QAO, the timing check and the Al sphere measurements may be repeated at the end of the survey day. These check requirements were based on data rates and the historical stability and reproducibility of each sensor type.

5.4.2 Emplaced Sensor Calibration Items

A calibration strip comprised of two replicates of each item of interest has been emplaced on site to verify proper system operation on a daily basis. The calibration strip will be surveyed each morning and each evening that data is collected. The data will be preprocessed, checked for data quality, and then the signal strengths and noise levels will be compared to the site-specific response curves and background levels to verify consistency of system performance. The planned schedule for the items of interest placed in the calibration strip is given in Table 3.

Table 3. Tentative Former Camp SLO Calibration Lane Configuration

Item	Depth	Orientation
60mm Mortar	3x diameter 4x diameter	EW EW
81mm Mortar	3x diameter 4x diameter	EW EW
4.2-in Mortar	3x diameter 4x diameter	EW EW
2.36" Rocket	3x diameter 4x diameter	EW EW
4-inch diameter ferrous sphere	3x diameter 5x diameter	

5.5 DATA COLLECTION PROCEDURES

5.5.1 Scale of Demonstration

The EM61MTADS arrays conducted total coverage surveys of the 11.1-acre demonstration site at the Former Camp San Luis Obispo. Threshold exceedances will be identified from each data set using an aggregate threshold determined from the response curves for each system and the items of interest. The site background levels and the depth of interest will also be included in the threshold determinations. A data segment around each threshold exceedance will be extracted,

analyzed, and dipole model fit parameters extracted. These results will be provided to the ESTCP Program Office in addition to the archival data.

5.5.2 Sample Density

EM61MTADS data are collected with nominal down-track spacing of 15 cm and cross track spacing of 50 cm. Because the three transmitters in the EM61MTADS array are synchronized, data are collected in two orthogonal directions to increase the number of "looks" or directions of illumination of each anomaly by the array. This effectively doubles the data density.

5.5.3 Quality Checks

Preventative maintenance inspections are conducted at least once a day by all team members, focusing particularly on the tow vehicle and sensor trailer. Any deficiencies are addressed according to the severity of the deficiency. Parts, tools, and materials for many maintenance scenarios are available in the system spares inventory which will be on site. Status on any breakdowns/failures which will result in long-term delays in operations will be immediately reported to the ESTCP Program Office.

For location data, the RTK GPS receivers present a Fix Quality value that relates to the quality / precision of the reported position. A Fix Quality (FQ) value of 3 (RTK Fixed) is the best accuracy (typically 3-5 cm or better). A FQ value of 2 (RTK Float) indicates that the highest level of RTK has not be reached yet and location accuracy can be degraded to as poor as ~1 m. FQs 1 & 4 correspond to the Autonomous and DGPS operational modes, respectively. Data collected under FQ 3 and FQ 2 (at the discretion of the data analyst) are retained. Any other data are deemed unsatisfactory, flagged, and not processed further. The section of data containing the flagged data will be logged for future re-acquisition as required. Data which meet these standards are of the quality typical of the MTADS system. Any data set which has been deemed unsatisfactory by the data analyst is flagged and not processed further. The area corresponding to the flagged data will be logged for future reacquisition. Data which meet these standards are of the quality typical of the MTADS system. For the EM61MTADS array, similar procedures are used, different only in the specific details of the data collected.

5.5.4 Data Handling

Data are stored electronically as collected on the MTADS vehicle data acquisition computer hard drives. Approximately every two survey hours, the collected data are copied onto removable media and transferred to the data analyst for QC/analysis. The data are moved onto the data analyst's computer and the media is recycled. Raw data and analysis results are backed up from the data analyst's computer to optical media (CD-R or DVD-R) or external hard disks daily.

These results are archived on an internal file server at NRL or SAIC at the end of the survey.

5.6 VALIDATION

At the conclusion of data collection activities, all anomalies on the master anomaly list assembled by the Program Office were excavated. Each item encountered was identified, photographed, its depth measured, its location determined using cm-level GPS, and the item removed if possible. All non-hazardous items were saved for later in-air measurements as appropriate. This ground-truth information, once released, was used to validate the objectives listed of the overall program and this project.

6 DATA ANALYSIS AND PRODUCTS

6.1 INTRODUCTION

This section steps through our data analysis and products in the order they were performed. We begin with a description of the data and then proceed step-by-step through our two iterations of modeling/risk-analysis/stop-digging threshold analysis. In summary, those steps may be described as follows:

- Description of the data
- Target polygon definition
- Polygon-based cannot analyze definition
- Remove non-target background noise
- Ellipse definition
- Attribute extraction
- Amplitude Discriminator
- Attribute Reduction
- LGP Modeling
- Risk Analysis
- Prepare Prioritized Dig List
- Request further ground-truth
- Iteration two Amplitude discriminator
- Iteration two Attribute reduction
- Iteration two LGP Modeling
- Iteration two Risk analysis
- Iteration two Prepare prioritized dig list

6.2 DESCRIPTION OF DATA

The data we received for analysis was of two types: DGM and target location together with partial ground-truth (for the training data only).

We used the same DGM in iteration one and iteration two. It was comprised of four channels of data for about 1.1 million spatially located data points from the EM61MTADS survey of the SLO site.

The target list or spatial information did vary from iteration to iteration.

For iteration one, we received spatial coordinates for 1464 targets identified by the program office and ground-truth for the training targets. Altogether, the 1464 targets were comprised of:

- 182 training (or “labeled”) targets. These were the Targets for which we knew ground-truth; and
- 1282 blind-data (or “unlabeled”) targets. These were targets for which we did not know ground-truth.

We augmented the labels provided to us by reviewing pictures of the training targets and assigning some items that were marked as Not-UXO by the program office as UXO. Our reasoning was that we were looking for relatively cylindrical, bullet shaped objects, not necessarily the particular ordnance items included in the training data. Accordingly, we marked cylindrical items greater than about 10 cm in length as “RML-UXO” and Table 4 shows the targets that were assigned a “UXO” label by this process.

Table 4. Targets assigned "UXO" label in iteration one

Target ID
468
549
562
579
1025
1322

The resulting groundtruth for the training data for the first iteration was as follows:

Table 5. Training data summary for first iteration

Type	Count	%
Not-UXO	148	81.32%
60mm Mortars	14	7.70%
81mm Mortars	3	1.65%
2.36" Rockets	7	3.85%
4.2" Mortars	4	2.20%
RML UXO	6	3.30%
Total	182	100%

For iteration two, we used the same spatial coordinates for the 1,464 targets identified above. However, the distribution of training and blind data was different. There, the 1464 targets were comprised of:

- 438 training (or “labeled”) targets. These were the Targets for which we knew ground-truth; and
- 1026 blind-data (or “unlabeled”) targets. These were targets for which we did not know ground-truth.

As in iteration one, the photos of the training targets were examined and as a result, additional Not-UXO were relabeled as UXO (“RML UXO”).

Table 6. Additional targets assigned a UXO label in iteration two

Target ID
376
475
525
637
797
945
953
964
1191

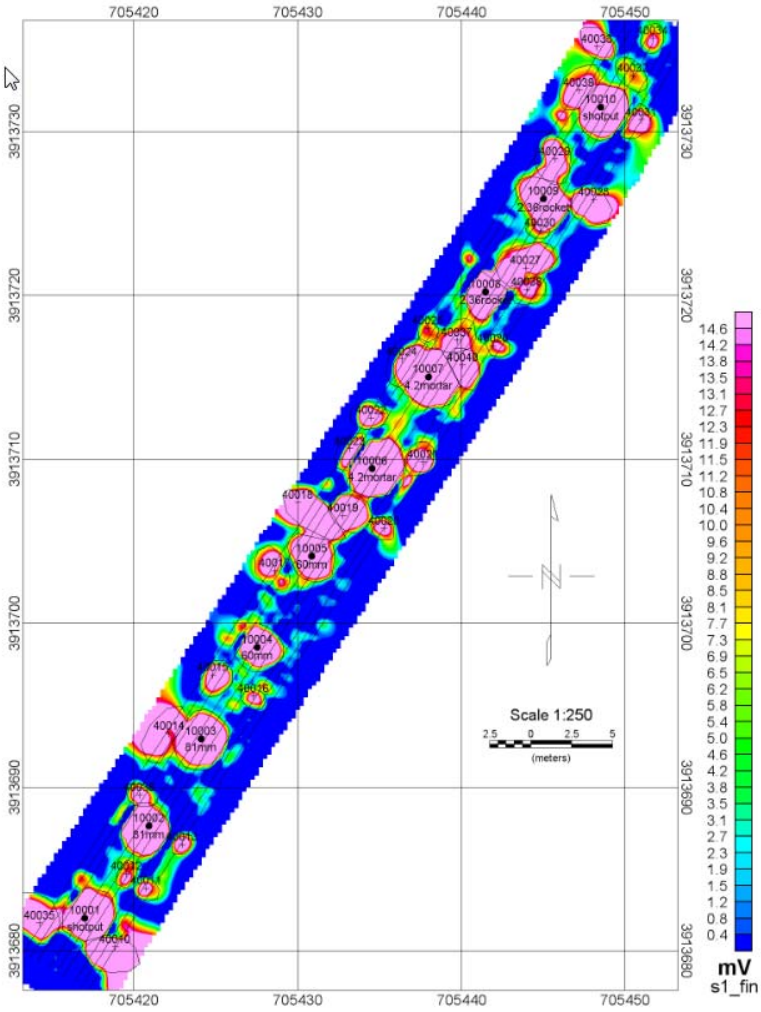
Once the reassignment of labels was performed, the ground-truth for the training data for iteration two consisted of the items described in Table 7:

Table 7. Training data ground-truth summary for second iteration

Type	Count	%
Not-UXO	316	72.15%
60mm Mortars	47	10.73%
81mm Mortars	24	5.48%
2.36" Rockets	11	2.51%
4.2" Mortars	24	5.48%
5" Rocket Warhead	1	0.23%
RML UXO	15	3.42%
Total	438	

We were also provided data for the calibration grid at SLO. Figure 10 shows the DGM for the calibration grid. From the perspective of this project, the important thing to note is the six black lines running through the calibration grid. These represent two parallel passes of the EM61MTADS array. By way of contrast, in the DGM for the SLO site, there were two sets of passes over each target, roughly perpendicular. So the data density for the site was about twice the data density of the calibration grid. We elected not to use the calibration grid data for training because the lower data density is likely to produce wider variance in the statistics produced for the calibration grid than the site itself. By itself, that would not be an issue. However, the calibration grid has a high proportion of UXO. That means the greater variance on the combined calibration data and training target data from the site would be correlated with the UXO/Not-UXO output we are trying to predict. The extent to which that would affect attribute reduction and modeling is not predictable. Accordingly, the calibration grid data were not used.

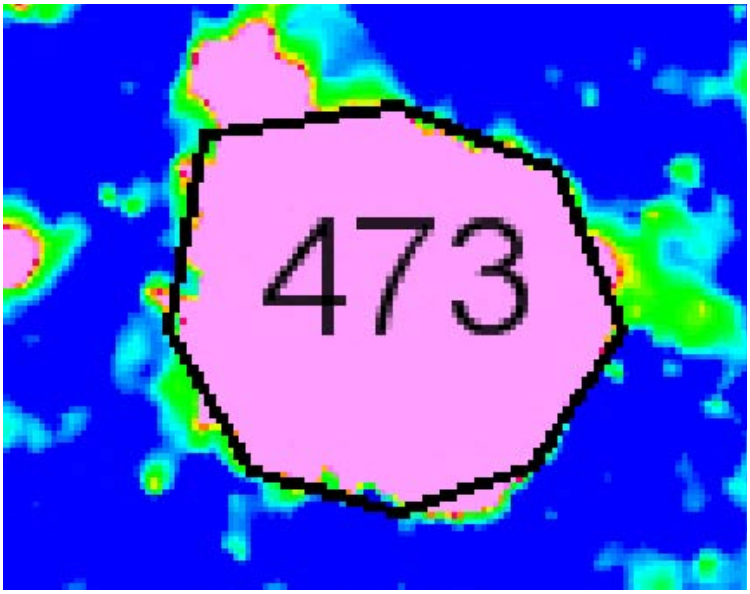
Figure 10. DGM of EM61MTADS on calibration strip at SLO



6.3 TARGET POLYGON DEFINITION

We defined a polygon for each program office target. This was done by visual inspection of a map produced on the data from Oasis Montaj using linear scaling, channel 1, a minimum millivolt setting of 0 and a maximum millivolt setting of 10. Figure 11 shows such a polygon.

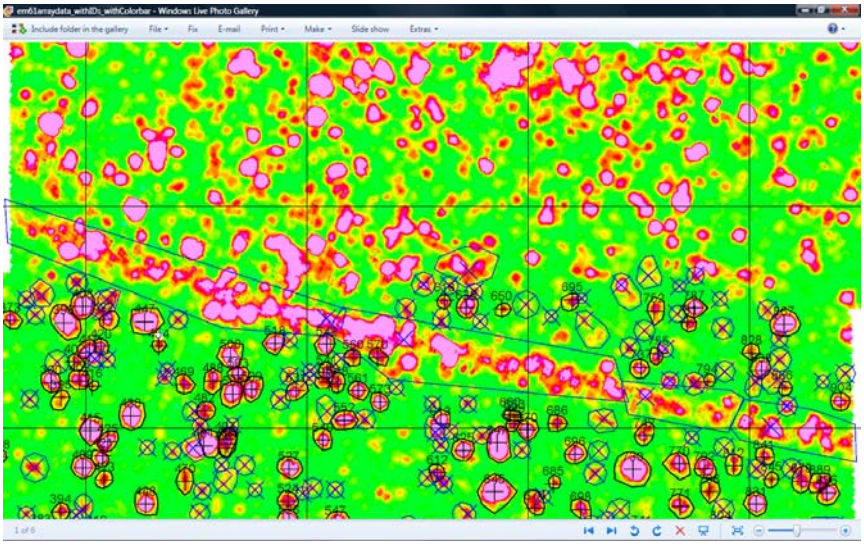
Figure 11. A target polygon



6.4 REMOVAL OF NON-TARGET BACKGROUND NOISE

Many anomalous regions in the Site were not designated by the program office as targets. We defined a polygon similar to Figure 11 for each such anomalous region using the same map settings as defined above. For irregularly shaped regions of this type, we removed the data points from our database. Figure 12 shows such a region. The pink linear region is defined by three polygons and those points are removed from the database.

Figure 12. Polygon enclosing irregular anomalous zone for which data points were removed from database



6.5 REMOVE CANNOT-ANALYZE CATEGORY ONE TARGETS

The next step was to identify targets for which good discrimination was not possible from visual examination of the polygons. As there are four points in this process where we identified cannot-analyze targets, the targets identified here will be referred to as the "cannot-analyze one" category of targets.

Cannot-analyze one targets due to overlap were identified by visual inspection from a map produced on the data from Oasis Montaj using linear scaling, Channel 1, a minimum millivolt setting of -50 and a maximum millivolt setting of 100. That map showed the polygons described in Sections 6.3 and 6.4.

On visual inspection, we identified four different criteria to assign targets to cannot-analyze one were as follows: (1) Overlapping Targets; (2) Targets with missing sections of DGM; (3) Targets with Local Data Inconsistency; and (4) Targets with Insufficient DGM Density to Support a Conclusion.[19]

We will discuss each criterion separately.

6.5.1 Overlapping Targets

Cannot-analyze one targets identified as a result of overlap were identified as targets where the extent and/or nature of the overlap would create ambiguity in the attributes extracted from the target. These criteria necessarily result from our process, which requires us to define an ellipse for every target. To do so, we first defined a polygon around the target and then fit an ellipse to the polygon. This ellipse definition process can fail if the overlap makes the boundary of the target ambiguous or, even if the boundary is reasonably clear, the extent or nature of the overlap make it unlikely that reliable attributes can be extracted from the target.

Overlapping targets on the Site naturally broke down into four different categories: (1) Blobs; (2) Distinct targets with too much overlap; (3) Ambiguous double-peaked/overlapping targets; and (4) Lower amplitude target near higher amplitude Target. We found, as a practical matter, that breaking the overlap issue out in this manner was useful. Many cannot-analyze one targets would have been excluded under multiple criteria. In our process, we excluded a target once any one of the cannot-analyze one criteria was met for that target.

A note is necessary here regarding the marking of program office targets and the marking of regions of interest that are NOT program office targets.

In the remainder of the maps in this "cannot-analyze one" section, program office targets are marked with a cross at the location selected by the program office (┼).

Anomalous regions that are not program office targets (see Section 6.4) are marked with an "X". For example, in the case of Figure 13, the "X" marks the center of a blob that was designated for the purpose of removing all targets and data points for subsequent analysis because the contents of the blob were designated as cannot-analyze one. In other figures, such as Figure 15, the "X" designates the center of an anomalous region that was marked to remove it from the background noise in the vicinity of program office targets (see Section 6.4) (the anomalous region in the lower right hand corner is an example of such a removal).

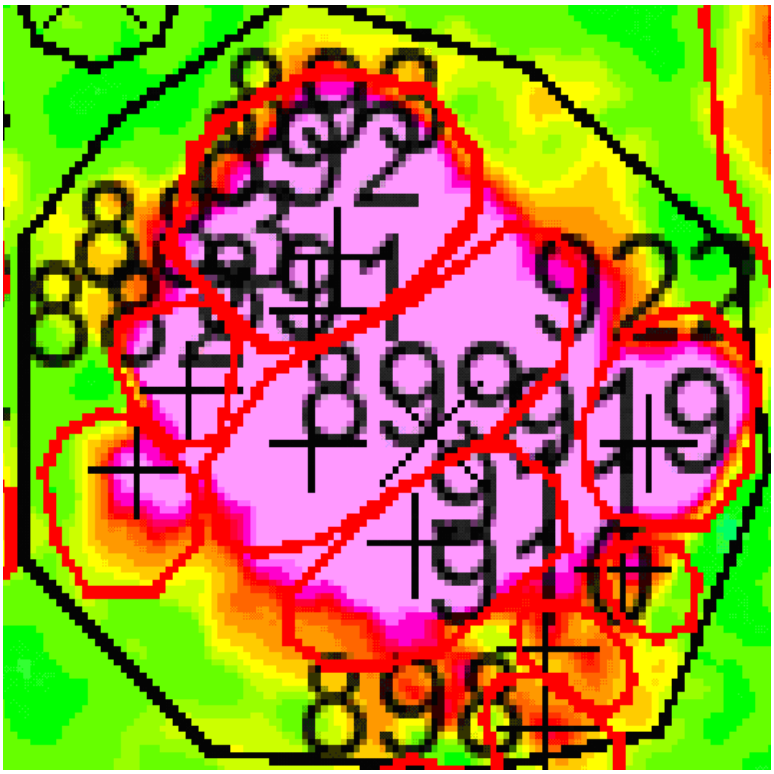
When numbers are shown on these maps, they represent program office master id's. The numbers appear above and centered on the marker ("┼") that marks the location of the program office target.

[19] These criteria were applied aggressively given the small number of examples of UXO in the training data set. We had hoped to relax our constraints somewhat in iteration two. But time did not permit.

6.5.1.1 Blobs

"Blobs" were identified by visual inspection. The Site contains many areas in which the density of targets selected by the Program Office is so high that it is not possible to define a polygon around each Target that separates the Target from either background noise and/or nearby targets in a reasonable manner. We identified those areas as blobs.

Figure 13. Example of a cannot-analyze one blob



In Figure 13, nine targets are tightly spaced. The red polygons represent our attempt to separate them from each other, in our judgment unsuccessfully. The black polygon shows the region containing targets we assessed for this reason as "cannot-analyze one."

Figure 14. Second example of a cannot-analyze one blob

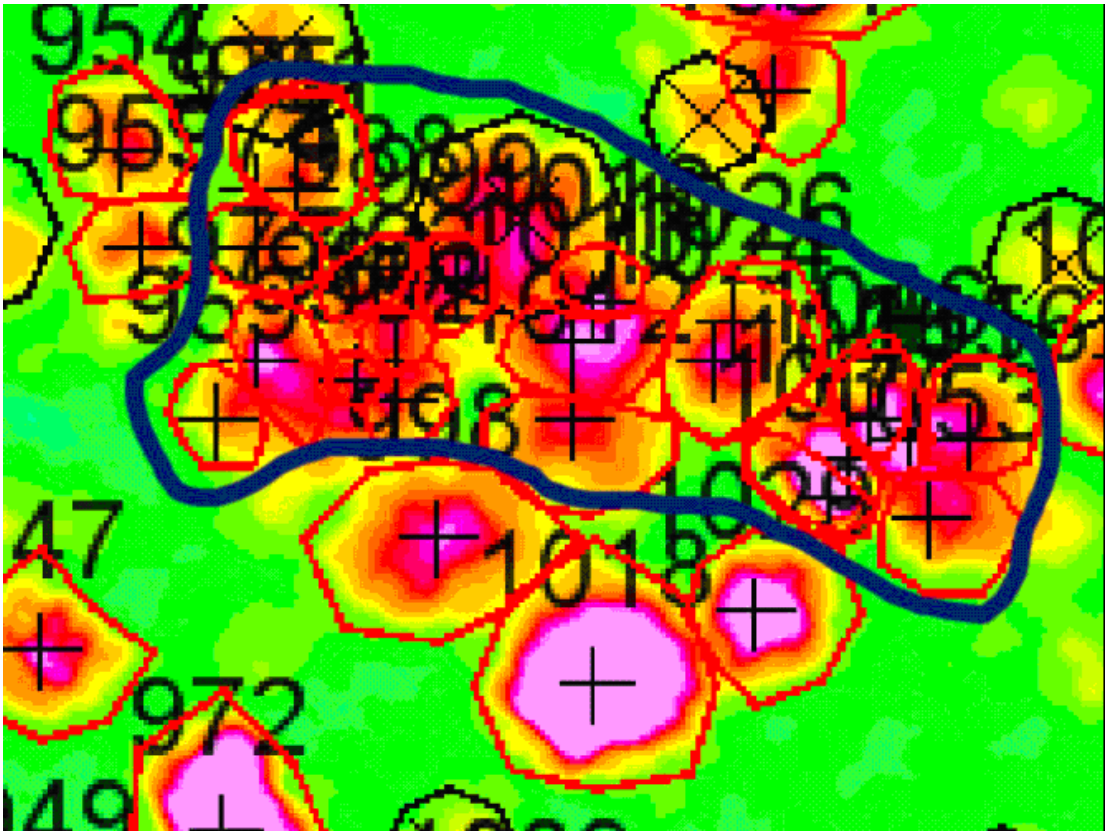


Figure 14 shows a second example of a more complex blob. Here, a simple polygon would encompass the three targets just below the irregular blue polygon. In this case, we created the complex polygon containing 21 targets to remove these blob targets and the points included in them from the analysis of the three targets below the polygon.

6.5.1.2 Distinct Targets with Too Much Overlap

The principle for this subcategory is similar to the blobs—the extent of overlap is such that we cannot determine appropriate boundaries for the polygons or too much of the adjacent targets falls in the overlap region to expect a good attribute set. Again, these determinations were made by visual examination of the above referenced map.

Figure 15. Three targets where we could not determine nature or extent of overlap

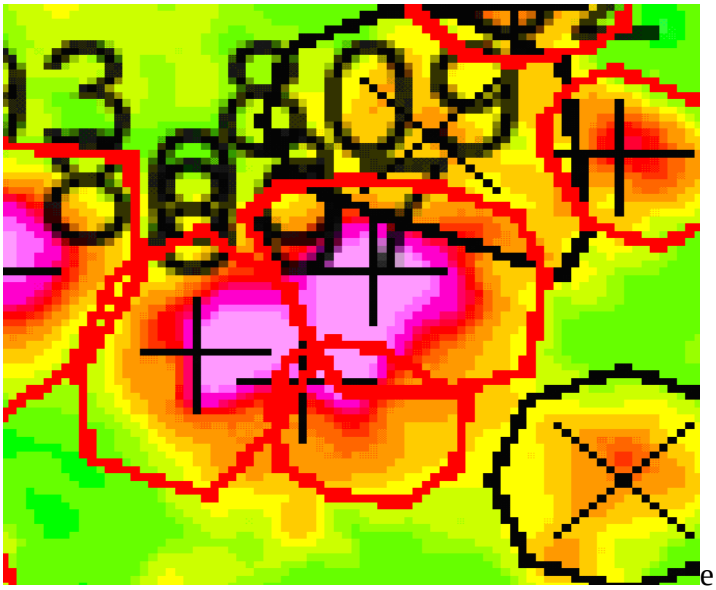
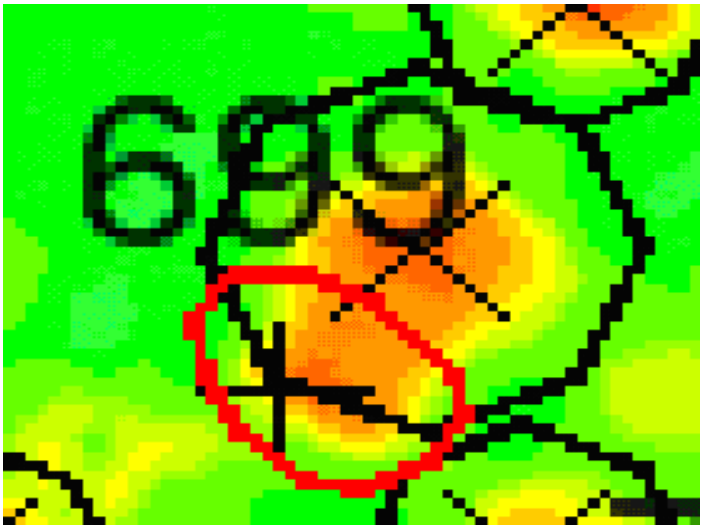
 e

Figure 15 shows three overlapping targets (the center three targets). While we attempted to draw polygons, visual examination makes it obvious that it is not possible to determine the proper extent of the bottom polygon and, therefore, the extent of the overlap between that polygon and the two targets above it. Accordingly, they were assigned to cannot-analyze one.

Figure 16. Overlapping nearby targets



Target 699 in Figure 16 was initially defined by the red polygon. The nearby anomaly up and to the right was not marked by the program office as a target, but was defined by us to remove background noise. On visual inspection, we determined that the extent and ill-defined boundary of the overlap was sufficient to warrant excluding target 699 as cannot-analyze one.

6.5.1.3 Smaller Target Overlapping Nearby Larger Target

This criterion is a variant on overlapping targets. The shoulders of a larger target beside smaller targets have the potential to affect the signal in the smaller target in an unpredictable way. In that situation, we assign the smaller targets to cannot-analyze one.

Figure 17. Several smaller targets surrounding a large anomaly

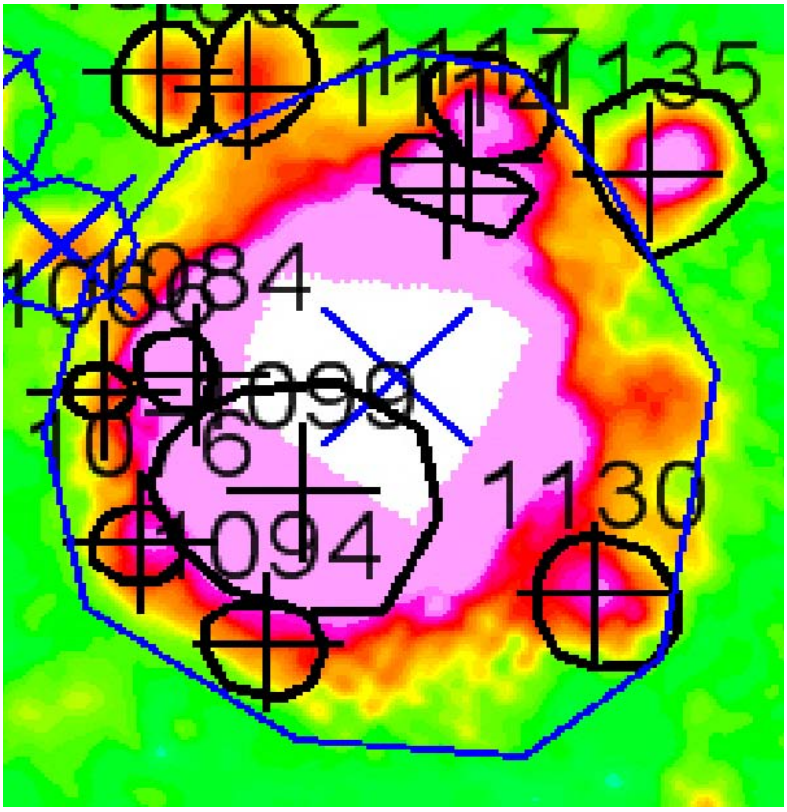


Figure 17 shows an example of this situation. The large white area contains no DGM. But an item or items in the white area obviously causes a large anomaly signature around it. That anomaly signature interacts with many of the smaller, nearby targets in an unpredictable manner, making the boundaries of the targets and attributes extracted unreliable. Thus, the program office picks around it cannot be reliably discriminated because of the unpredictable effect of the larger target on them.[20] By way of contrast,

Target 1135 in Figure 17 (upper right hand corner) has distinct boundaries with the background noise region except for a small overlap with the large anomalous region. So Target 1135 was not assigned to cannot-analyze one by this criterion.

6.5.1.4 Ambiguous Double-Peaked/Overlapping Targets

This issue is related to the overlapping targets issue. The munitions on this site contain a relatively large number of long cylindrical items. The EM61MTADS will frequently produce a

[20] For reference, the large blue x and the surrounding blue polygon were placed by us to designate an entire region that was removed from subsequent analysis both as targets subject to analysis and as signal to be used to analyze surrounding targets.

double peak when it traverses such items. Accordingly, ambiguities arise as to whether two peaks are two distinct targets or whether they are a single target with two peaks. This section describes the rule we applied to handle that ambiguity. The important thing in applying such a rule is that it be applied consistently across the training and blind data so that our learning algorithms will be learning to discriminate based on consistent data.

The target selection process by the program office addressed this issue in the following manner: under each peak that exceeded the threshold, an inversion was run. In addition, the inversion was run under adjoining peaks. If the inversions for a single target area containing two peaks showed only one object in the area with high coherence and sensible parameters, then it was identified as a single target. If the inversions identified more than one object in that area with high coherence and sensible parameters, it was identified as multiple targets. If the inversions were ambiguous, the determination was made manually by the program office.

The program office did not identify which of these targets were set by high-coherence inversions and which were set manually. Accordingly, we could not count on a particular target pick as being the result of a high-coherence/sensible parameter inversion and therefore developed a consistent rule to apply to training and blind data to accommodate this uncertainty.

Our procedure was as follows:

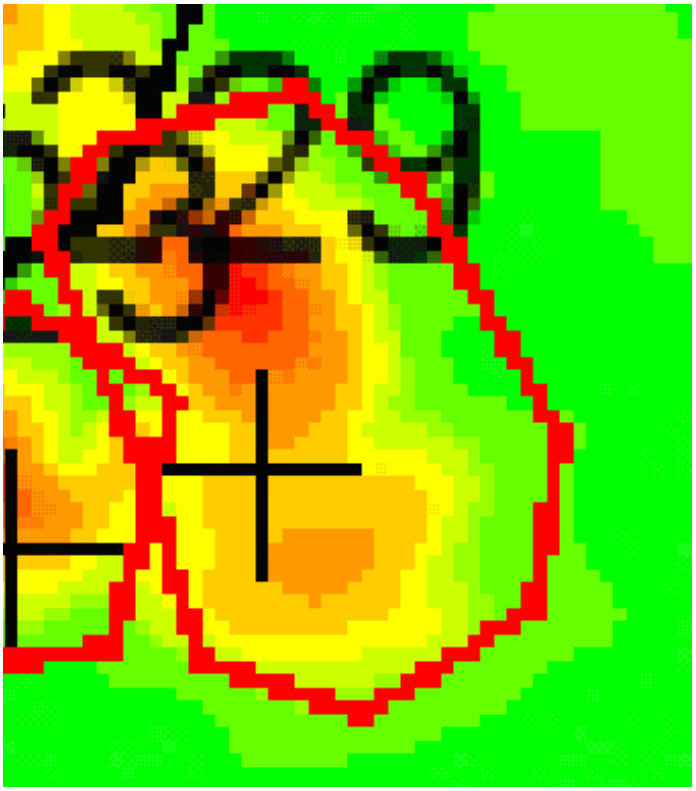
The SAIC portion of the team drew polygons for all program office targets, including the potential double peaked targets. Assisted by inversions, they determined whether the overlapping adjacent targets were most likely one or two targets. The targets were not, however changed as a result of this process.

The RML portion of the team then took the program office determinations and the SAIC polygon determinations and treated them as two expert opinions on the nature of the target or targets. We then applied the following decision rules to these expert opinions in the following situations:

1. The Program Office and the SAIC portion of the team both assessed the region as a single target. That judgment was accepted and the target was not assigned to cannot-analyze one.
2. The Program Office assessed the region as two separate targets and the SAIC portion of the team assessed it as one. The “Two Targets Rule” below was applied.
3. The Program Office assessed the region as one target and the SAIC portion of the team assessed the region as two. The “Two Targets Rule” below was applied.
4. Both the Program office and the SAIC portion of the team assessed the region as two targets. The “Two Targets Rule” below was applied.

Figure 18 illustrates a situation where both the Program Office and SAIC assessed the region as a single target:

Figure 18. Rule 1. Target NOT assigned to cannot-analyze one



The “Two Targets Rule” referred to above was applied whenever either the program office or SAIC designated a region as two targets. The rule was as follows: If the program office picked location for either of the two targets was noticeably toward the boundary between the two targets, and away from the apparent peak of the target on the map, we assigned the target to cannot-analyze one. Otherwise the two targets were not assigned to cannot-analyze one.

Figure 19 illustrates the application of the Two Targets Rule in the situation where the Program Office designated a region as one target but SAIC designated it as two (see target 702—red polygon center left and the black polygon just to its right). Because the program office pick location (the cross) tends strongly toward the boundary region, target 702 was designated as cannot-analyze one.

Figure 19. Illustration of Rule 2. Target assigned to cannot-analyze one because of two-targets rule

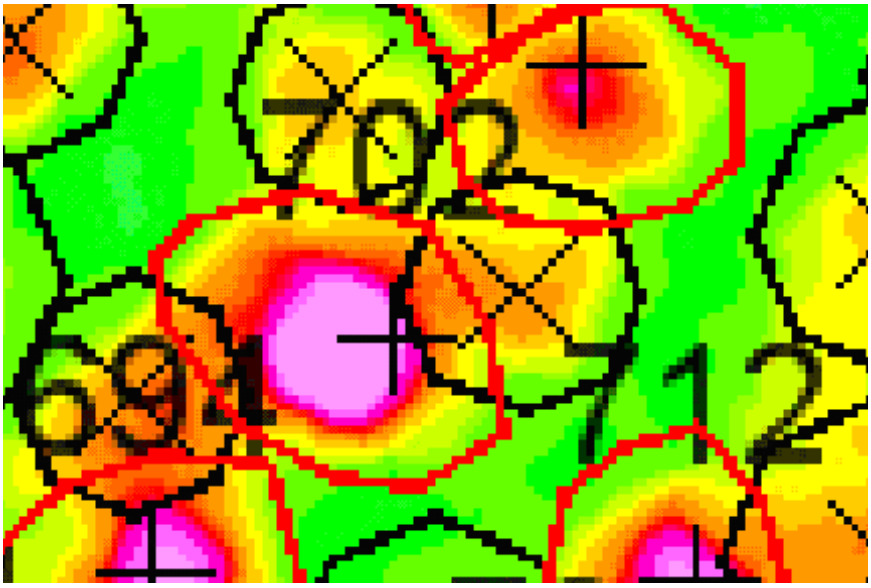


Figure 20 and Figure 21 show the application of this Two Target Rule to assign targets to cannot-analyze one in the situation where both the Program Office and SAIC assessed the region as containing two targets. In both figures, the location of one of the program office target locations tends strongly toward the overlap region between the targets. We take that as evidence of ambiguity in the target assessment, possibly by an inversion that places a target location where it would be expected for a double peak target and the targets are assigned to cannot-analyze one.

Figure 20. Overlapping target assigned to cannot-analyze one because of two targets rule

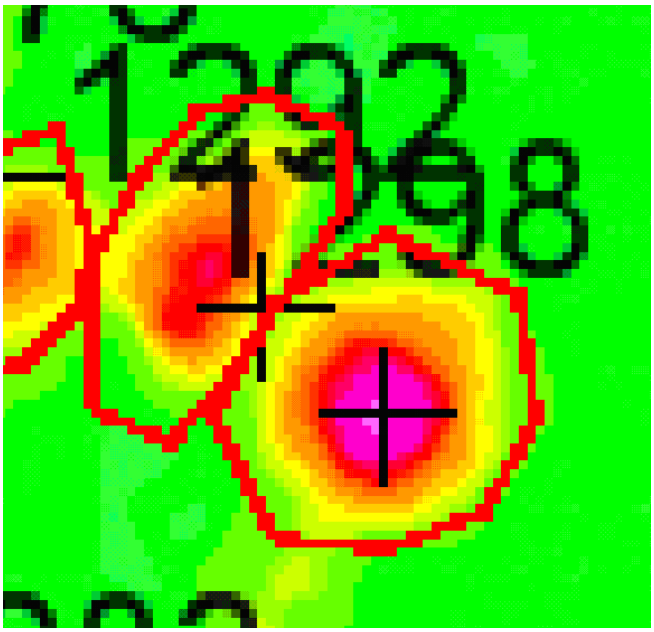
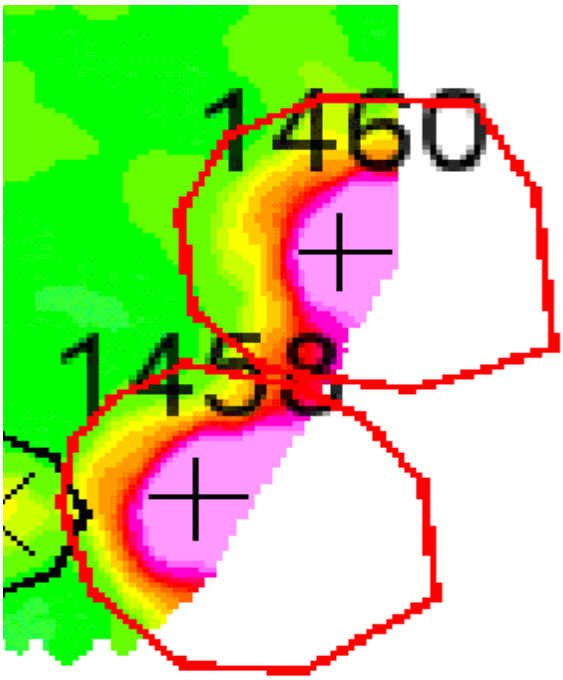


Figure 21. Overlapping target assigned to cannot-analyze one because of two targets rule

6.5.2 Targets with Missing Sections of DGM

Some targets did not have sufficient DGM to confidently define the target boundaries. These were assessed as cannot-analyze one. Figure 22 is an example of two such targets.

Figure 22. Targets with missing DGM

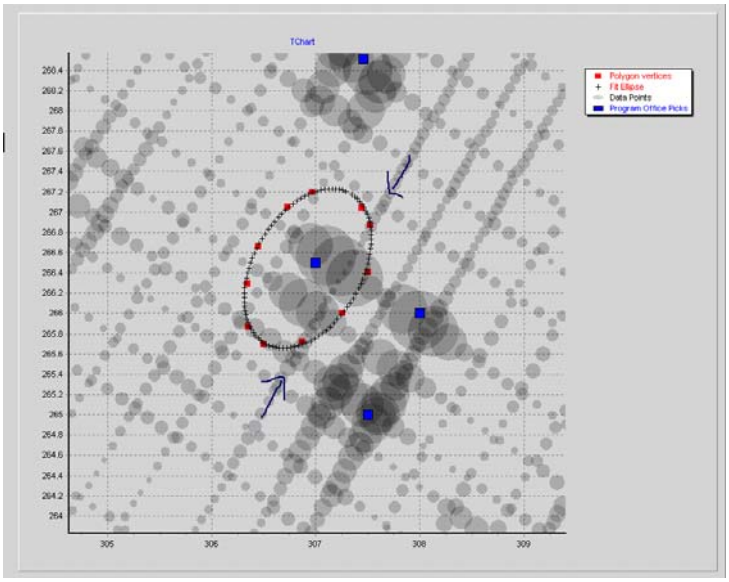


6.5.3 Local Data Inconsistency

We used visual inspection of a bubble chart to assess local data inconsistency. Local data inconsistency is a line of data in a target region that is clearly inconsistent with the remainder of that data. No fixed parameters were used for this analysis because each target required that we set the maximum millivolt reading to expose the details of that particular target.

Figure 23 shows a target manifesting local data inconsistency.[21] Each bubble represents a single DGM reading for a single channel. The size of the point is linearly proportional to the millivolt reading for that point. The program office picks are represented by blue squares. Note the substantial inconsistency between the line of data highlighted by the two arrows and the remainder of the data in the target.

Figure 23. Example of local data inconsistency. Scale on both axes is meters.



6.5.4 Cannot-Analyze One Results

Using the rules listed above 235 (25 training and 210 blind) targets were classified as cannot-analyze one. Table 8 shows the sample size before and after the cannot-analyze one targets were removed.

Table 8. Cannot-Analyze One

Iteration 1	Train	Blind	Total
Original Data Sample Size	182	1282	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	157	1072	1229

[21] In Figure 23 the scaling was linear, the minimum millivolt setting was -25 and the maximum millivolt setting was 100.

6.6 ELLIPSE DEFINITION

The ellipses for our targets and non-target anomaly regions were fit to the target polygons using least mean squared error (i.e. the fit between the polygon vertices and the ellipse) as the criteria. We then converted the polygons into ellipses. The process for that was straightforward. We used a downhill simplex optimizer to find the ellipse that minimized the mean squared error between the vertices of the polygon and the ellipse.

Table 9. Output of downhill simplex fit of ellipse to manually defined polygon--targets 1-34

Data_Id	TId	a	b	x	y	theta	MSE	Niters
27004	1	2.4684405671...	2.1331004341...	120.2447521161	339.26511823...	0.4045799985...	0.00616001941250834	441
27005	2	0.8155569097...	0.7484379437...	120.18481339...	354.38611335...	0.4738706252...	0.00095070459443789 3	501
27006	3	1.3228275908...	1.1160407023...	120.29788027...	363.74165090...	0.3245952205...	0.00108675574152634	501
27007	4	1.7945551220...	1.0980378712...	120.07658586...	369.46942566...	-0.1458064599...	0.00140352367904032	498
27008	5	2.2258878166...	1.5752547591...	120.18635215...	325.58795386...	-1.2979229815...	0.0148579160282459	399
27009	6	1.0996345715...	0.7648490680...	121.25904797...	321.14499197...	0.7855823855...	0.00189584977717347	501
27010	7	1.2889058097...	0.9174878247...	121.29346637...	361.54827001...	-0.6977175759...	0.00240441972051258	461
27011	8	0.9353209009...	0.5252346604...	122.06555912...	321.98477321...	0.7973052876...	0.000683710109176272	501
27012	9	1.1747736492...	1.1747736492...	121.95645925...	327.02367649...	0	0.000683710109176272	501
27013	10	0.9072413337...	0.7998062102...	122.18267910...	370.03367393...	0.3203147999...	0.000801172525917344	501
27014	11	2.2703176527...	1.5368942357...	122.12908470...	334.93231586...	0.7018144952...	0.0200777637096876	501
27015	12	2.0089343919...	1.5992558385...	122.34785225...	359.47568395...	-1.0117479288...	0.0149305913968729	337
27016	13	1.6950414920...	1.3912658711...	123.08568497...	328.39718808...	0.1055840980...	0.00436383977699084	415
27017	14	2.3924061326...	1.4723430369...	123.88415942...	323.95854937...	-1.3332473325...	0.00519709560050544	429
27018	15	1.1863306246...	0.9662592786...	124.20253320...	354.04195293...	1.5584108876...	0.00161889272068589	393
27019	16	1.3505296696...	1.2154069594...	125.44342305...	366.72991446...	0.3332250910...	0.00637706001686258	486
27020	17	1.4613464537...	1.3196015143...	127.35971469...	338.27016386...	-0.7152954878...	0.00197967789699234	485
27021	18	0.8800300547...	0.7765898030...	127.50221758...	328.29312825...	-0.9428677623...	0.00773532389113423	501
27022	19	1.5735990306...	1.3156596906...	128.64992256...	361.05128297...	0.0135566602...	0.017308369868784	489
27023	20	2.0543104630...	1.8773825336...	128.66272600...	335.40446768...	-1.1982799078...	0.0272779534734312	402
27024	21	2.0786327860...	1.9319007798...	130.81691897...	344.31820998...	-0.6376443070...	0.0352377216763718	501
27025	22	1.5144547904...	1.0291620510...	130.55188082...	319.21075458...	0.3444270963...	0.000660151379785944	351
27026	23	1.5397714896...	1.1691407046...	132.75659135...	359.25075227...	1.2837552660...	0.016242468393665	484
27027	24	2.0260097472...	1.6608740137...	133.33341495...	339.90726439...	1.5043142104...	0.00979313046854225	396

The "TId" column in Table 9 is the program office's Master ID. The extracted parameters of the ellipse are:

- "a": The semi-major axis of the ellipse in zeroed meters;
- "b": The semi-minor axis of the ellipse in zeroed meters;
- "x": The X coordinate of the ellipse in zeroed meters;
- "y": The Y coordinate of the ellipse in zeroed meters;
- "theta": The rotation of the ellipse in radians. Rotation is counterclockwise from an x-axis orientation.

The MSE column shows the mean squared error of the fit between the polygon vertices and the fit ellipse. These targets show a very close fit between the ellipse and the polygon.

6.7 ATTRIBUTE EXTRACTION

Once the ellipses had been defined, we removed data points from the ellipse that were in the ellipse of an adjacent target or non-target anomalous region. Attributes are extracted from the remaining points in the target ellipse.

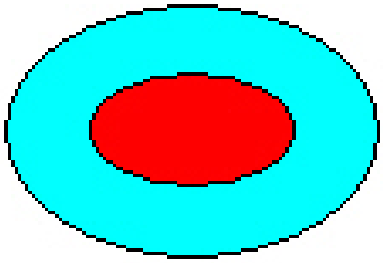
Attribute extraction is the process of converting the DGM in the vicinity of a picked target into meaningful statistics about the target. For this project, we extracted and used three types of attributes:

- Attributes that measure a statistic of the amplitude of the signal value of a single channel ("Amplitude Statistics); and

- Attributes that measure the ratio as between two different channels of Amplitude Statistics ("Ratio Statistics").
- Attributes that measure the ratio of adjacent Ratio Statistics ("Rate of Change Statistics").

Attributes were calculated on the DGM data points within different regions around the target. Figure 24 illustrates those regions. The ellipse in that figure is the entire ellipse defined as above around the target. The red and blue regions are sub regions in the ellipse from which features are extracted from the DGM data points contained therein.

Figure 24. A simple illustration of ellipsoidal rings for attribute extraction



The attributes calculated, for each target, consisted of the first three moments. Each of these three moments was calculated for each of the different regions around the target, including the entire ellipse and the two sub-regions as follows:

5. For Amplitude Attributes: The value for channels 1, 2, 3, 4, and sum;
6. For Ratio Attributes: The values for all possible ratios between the DGM value for channels 1,2,3, and 4.
7. For Rate of Change Attributes: The value of all Ratio attributes, respecting the decay order of the channels (e.g. Ratio of Channel 1 to Channel 2 / Ratio of Channel 2 to Channel 3).
8. The result of this process is hundreds of attributes for each target. They are databased and used for subsequent analysis.

6.8 REMOVE CANNOT-ANALYZE TWO CATEGORY TARGETS

At this point, we determined whether there were insufficient data points in any region for any target to exclude that target from later analysis. If the sample size of data points, for any target, was less than or equal to 5 on any of the Amplitude Statistics in the target ellipse, then it was excluded as a "cannot-analyze two" category target due to insufficient Amplitude DGM density. Table 10 shows the adjustments we made to the cannot-analyze list and our remaining targets as a result of this analysis.

Table 10. Targets lost due to cannot-analyze two analysis

Iteration 1	Train	Blind	Total
Original Data Sample Size	182	1282	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	157	1072	1229
Cannot-Analyze 2 (Low Sample Size)	(1)	(1)	(2)
Sample Size After Cannot-Analyze 2 Removed	156	1071	1227

6.9 ITERATION ONE

We performed our first iteration using two modeling steps: (1) An amplitude discriminator; and (2) LGP modeling. Each step included a separate risk analysis. After completing these two steps, we produced a prioritized dig list. This section describes those iteration one steps.

6.9.1 Amplitude Discriminator

6.9.1.1 Introduction

We had found it useful in previous work to filter targets with an amplitude discriminator. An amplitude discriminator is a single attribute derived using Amplitude Statistics that, by itself, discriminates. That is, the derived attribute itself is the model.

The purpose of this was to use a very simple initial model to identify a set of targets that may be excluded as high-probability Not-UXO. The purpose of using only Amplitude Statistics is our observation that they are more robust against noise than the Ratio Statistics. Once excluded, the distribution of the Ratio Statistics stabilize and provide robust prediction.

6.9.1.2 Attribute Identification

Initially, we examined the extracted Amplitude Statistics to determine which were most predictive of the UXO/Not-UXO classification on the iteration one training set.

To do this, all of the extracted attributes were binned using a Chi-square binning procedure. Once binned, the count of UXO in the low and high bins were examined for each attribute to determine which had the lowest number of UXO in the low (or high) bin and the highest number of Not-UXO in the low (or high) bin.

Attribute AD was selected as the Amplitude Discriminator because it had 0 UXO and 51 Not-UXO targets in bin one. Table 11 shows the result of the Chi-square binning procedure for Attribute AD. Attribute AD may be described as follows:

- Attribute AD = Third moment for Channel 1 (first decay channel) for the entire ellipse.

Table 11 shows the result of binning the training data for with attribute AD.

Table 11. Bin information for attribute AD

Attribute AD	Bin Boundary	Bin Boundary	Total Targets	UXO Targets	Non-UXO Targets
Bin 1	-INF	0.265925565	51	0	51
Bin 2	0.265925565	INF	105	33	72

Chi square for the 2x2 contingency table of the last two columns of table with Yates correction is 16.48. The probability of that chi square with one degree of freedom is less than 0.0001. This is a statistically significant result.

6.9.1.3 Attribute Match Between Training and Blind Data

Because the training data will be used to make predictions to the blind data we identified how well the distributions of training and blind data "match". In looking at the descriptive statistics in Table 12 and the graphics in Figure 25 and Figure 26 it was determined that the training and blind data were consistent in their distributions.

Table 12. Descriptive statistics for attribute AD

Attribute AD											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1227	0.35049	0.23553	0.05547	0.3403	0.36055	1.36236	-0.377	0.000*	-0.225	>.10
Training Data	156	0.36417	0.23591	0.05565	0.37167	0.36714	1.09797	-0.333	0.087	-0.48	>.10
Blind Data	1071	0.3485	0.23552	0.05547	0.33711	0.36	1.36236	-0.384	0.000*	-0.186	>.10

Figure 25. Box and whisker plots for attribute AD

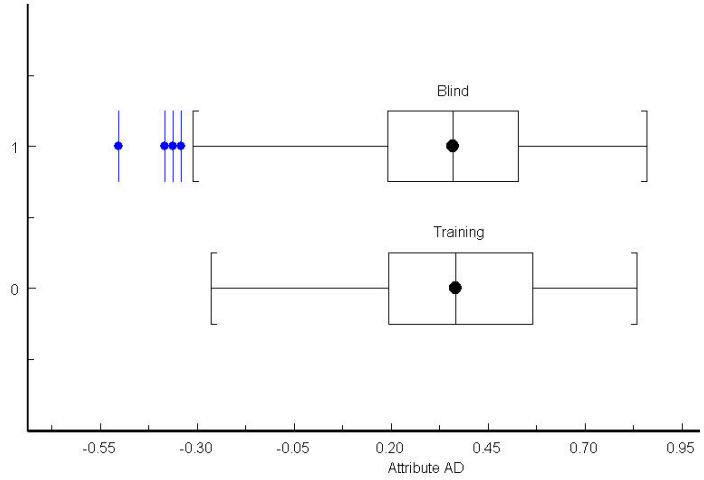
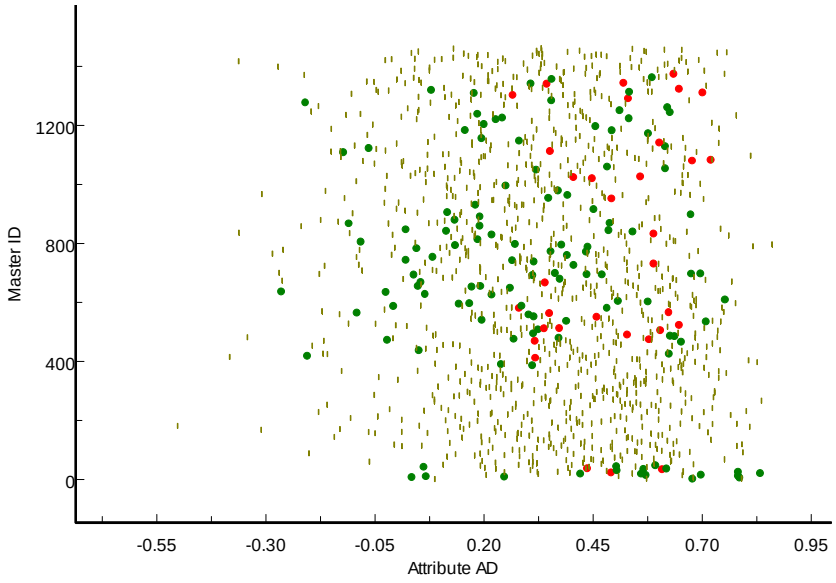


Figure 26. Attribute space for attribute AD (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data). Note, the y-axis is Master ID which is used to spread out the targets in the graph.



A note about Figure 26: Master ID is plotted on the y-axis. Obviously it has no meaning in terms of UXO vs. Not-UXO. However attribute space is one-dimensional (only Attribute AD). So the y-axis spreads the data out so it may be visualized better. The important point to take from it is that there are no UXO below AD=0.265. Thus Attribute AD by itself is a discriminator and we used it as such in the first iteration.

6.9.1.4 Risk Analysis/Stop-Digging Threshold

The next step was to determine which targets could be excluded as high-probability Not-UXO based on Attribute AD. In other words, were Attribute AD our only discriminator, where could we safely stop digging? The AD values were first converted into ranks across the entire training and blind data. Lower AD values were interpreted as larger rankings. Next, kernel regression with a Gaussian kernel was used to determine the probability of UXO for each target using the training targets:

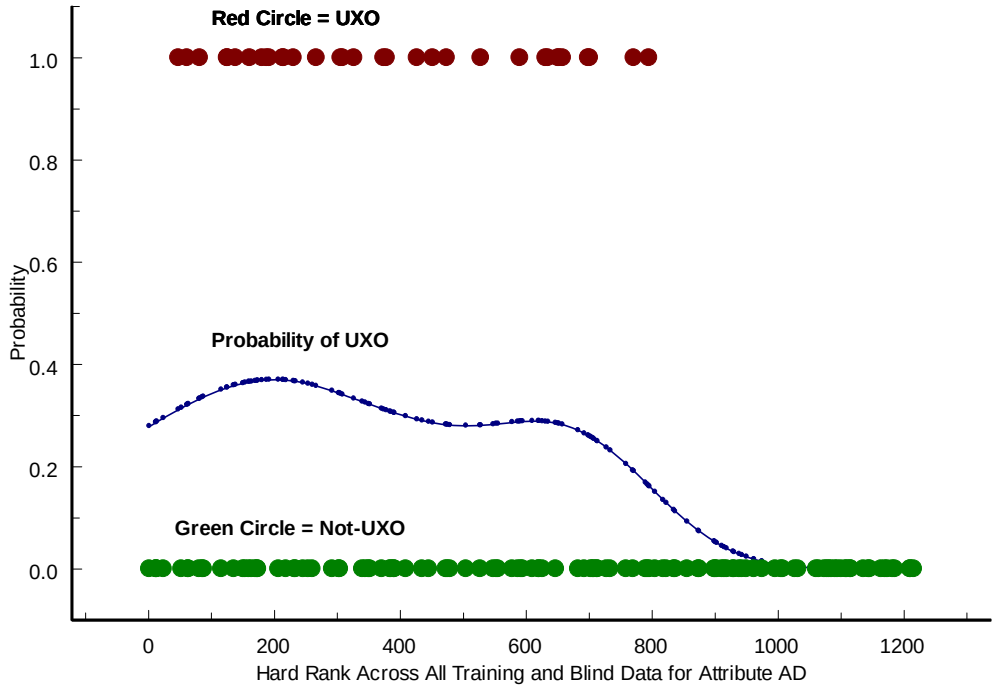
$$\mathrm{P}(UXO)_i = \sum_j e^{-\left(\frac{(x_i - x_j)^2}{2\alpha^2}\right)}$$

Where: (1) α represents the standard deviation of the Gaussian kernel; (2) x_i represents rank of the *ith* ranked blind data instance computed from the AD values across all training and blind data points; and (3) x_j represents rank of the *jth* ranked training data instance value of the AD values across all training and blind data points.

The value determined for the parameter, α, was 102.923. That value was determined by n-fold cross-validation on the training data. The α parameter selected was one that produced the minimal value for -2*log likelihood over the held-out training data from cross-validation, which is the maximum likelihood estimator for these data, assuming Bernoulli errors.

Figure 27 shows the derived model plotted against the rankings of the UXO and Not-UXO on the training data. Note that the rankings are derived from Attribute AD and represent the rankings across all training and blind targets not assigned to the cannot-analyze one or two categories.

Figure 27. Kernel regression fit between probability of UXO and Attribute AD on the training data

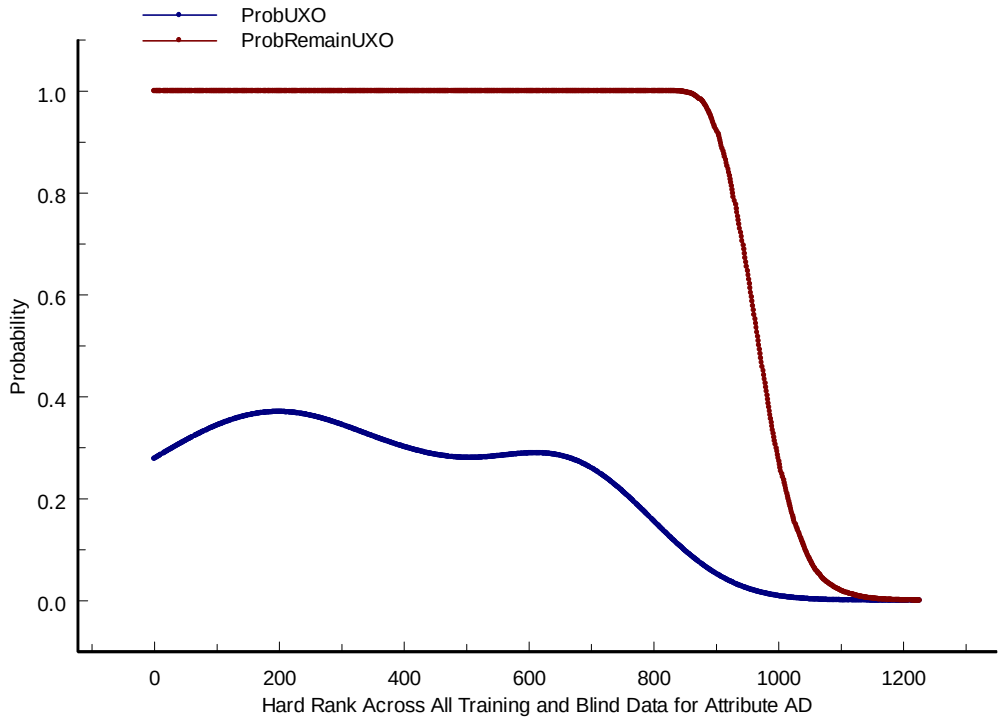


The Gaussian kernel, generated by the training data, using the above kernel width parameter, was then applied to the ranked blind data, generating a probability that each blind data item is UXO as a function of rank.

Once individual target probabilities are set, the probability that all blind targets above each AD ranking contain one or more UXO is calculated using the approach outlined in Section 2.1.6. This is the residual risk of UXO as a function of rank. In particular, we used Equation 1 and Equation 2 to compute the OR of the probabilities for all targets from the AD ranking for which the computation is being performed to the most extremely ranked blind target.

The blue line in Figure 28 is the probability of UXO on the blind targets as a function of the Attribute AD-based rank. The red line in that figure is the probability that one or more blind target UXO remains on site at each rank, assuming all targets to the left of that rank have been excavated. When the red line reaches a critical probability value ($p\text{-value}_{crit}$), we assess all targets remaining to the right of that rank (i.e. targets with a larger rank) as high-probability Not-UXO.

Figure 28. Kernel regression applied to blind data



A 95% confidence level was chosen at the start of this project and is used throughout this project. Since there were two risk analyses used to determine two stop-digging thresholds (Attribute AD and the LGP ensemble predictor described later) the Bonferonni correction should be applied.[22] Using the Bonferonni correction the $p\text{-value}_{crit}$ was set to .025 (2.5%).

The $p\text{-value}_{crit}$ was then used to determine the critical rank value of 1092, inferring that any target with a rank value greater than 1092 was high-probability Not-UXO. At that point, the probability of remaining UXO was 0.02480 — in other words, it satisfies the $p\text{-value}_{crit}$ criterion.

Table 13 shows the details for the 95% stop-digging threshold.

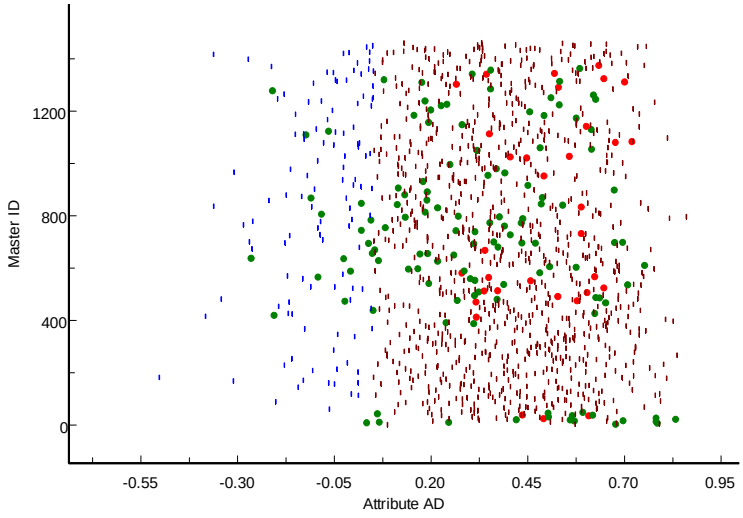
Table 13. Stop-digging threshold

Amplitude Discriminator						
Confidence	**TID**	**Score**	**HardRank**	**ProbUXO**	**ProbRemainUXO**	**% Of Blind Data Left in Ground**
95.00%	183	0.051282151	1092	0.00088	0.02480	9.20%

Figure 29 shows which blind targets were determined by this process to be safe to leave in the ground. Small blue lines represent the blind targets that would be left in the ground as high probability Not-UXO at this step. Brown represents blind targets may not be left in the ground—that is, they continue on for further analysis. Red and green circles represent UXO and Not-UXO in the training data, respectively. The x-axis shows Attribute AD while the y-axis shows Master ID. The Master ID is used only to separate the targets for better visualization.

[22] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

Figure 29. Attribute space for Attribute AD (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data above the stop-digging threshold. Blue Lines are blind data below the stop-digging threshold). Note the y-axis is Master ID, which is used to spread out the targets in the graph.



The result of this first iteration amplitude discriminator identified 118 blind targets and 17 training targets as high probability Not-UXO. After removing these targets from further analysis, there were 953 blind and 138 training targets remaining. Table 14 shows the updated sample sizes after taking into account the amplitude discriminator.

Table 14. Count of targets removed due to amplitude discriminator

Iteration 1	Train	Blind	Total
Original Data Sample Size	182	1282	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	157	1072	1229
Cannot-Analyze 2 (Low Sample Size)	(1)	(1)	(2)
Sample Size After Cannot-Analyze 2 Removed	156	1071	1227
Amplitude Discriminator Low Prob UXO	(17)	(118)	(135)
Sample Size After Amplitude Discriminator Removed	139	953	1092

6.9.2 Remove Cannot-Analyze Category Three Targets

Before continuing on to attribute reduction the sample sizes for each target that survived the attribute discriminator was calculated. If the sample size in any region (rings or entire ellipse) for a target, was less than or equal to 5 then that target was excluded as a cannot-analyze three category target.

Table 15 shows the updated sample sizes after removing the cannot-analyze category three targets.

Table 15. Count of targets lost due to cannot-analyze category three

Iteration 1	Train	Blind	Total
Original Data Sample Size	182	1282	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	157	1072	1229
Cannot-Analyze 2 (Low Sample Size)	(1)	(1)	(2)
Sample Size After Cannot-Analyze 2 Removed	156	1071	1227
Amplitude Discriminator Low Prob UXO	(17)	(118)	(135)
Sample Size After Amplitude Discriminator Removed	139	953	1092
Cannot-Analyze 3 (Low Sample Size - 2)	(1)	(10)	(11)
Sample Size After Cannot-Analyze 3 Removed	138	943	1081

6.9.3 Attribute Reduction for LGP Modeling

The Attribute Extraction process produces hundreds of statistics for every target. The goal in attribute reduction is to reduce the number of attributes used in modeling to just a handful of highly relevant attributes that contain complementary information content about the modeling problem.

6.9.3.1 Data Reduction Tools

We used a collection of tools at different points in the modeling process to obtain the critical few attributes.

The first step was to bin all of the attributes using either an equal-frequency or Chi-square binning procedure. Binning numeric variables is a fundamental technique in statistics and machine learning. Binning is the process of assigning numeric values to discrete categories. The two different binning procedures were used in this project were:

Equal-Frequency Binning. In equal-frequency binning, a number of bins is specified and the numeric values are divided into that number of bins. This technique attempts to assign the same number of numeric values to each bin. Sometimes that is not entirely possible because of tied numeric values.

Chi-Square Binning. Chi-square binning splits the numeric values into bins based on how well the splits do in minimizing the probability of Chi-square statistic of the 2x2 contingency table formed by the split of UXO and Not-UXO on either side of the bin boundary. This is a recursive technique. It starts by finding the single split that has the lowest probability. If the probability is greater than a selected parameter, binning stops. If it is less, then each bin is split in the same manner. Splitting continues recursively in each bin partition until the probability is greater than the set probability parameter.

Once binned, the mutual information between the attributes (independent variables) and UXO classification (dependent variable) is assessed. Mutual Information is usually one of the first measures we look at. It is used to get an idea of which attributes contain a high degree of information on whether or not a target maybe classified as a UXO. Formally, the mutual information of two discrete random variables X and Y may be defined as:

$$I(X;Y) = H(Y) - H(Y \mid X)$$

Where H(Y) is the marginal entropy of Y:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y)$$

and H(Y|X) is the conditional entropy of Y given X:

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x)$$

We will refer to mutual information between attribute X and UXO as *I*(X;UXO).

Mutual information favors attributes with more bins. This does not pose a problem for the equal-frequency binning since the number of bins for every attribute will be equal, but the Chi-square binning procedure will produce attributes will varying bin sizes. To correct for this, the symmetric uncertainty measure is used as an alternative for mutual information. Symmetric uncertainty is an expansion of mutual information that compensates for mutual information's bias toward attributes with a higher number of bins by dividing mutual information by the average entropies of X & Y. Formally, the symmetric uncertainty two discrete random variables *X* and *Y* may be defined as:

$$SU(X;Y) = 2 * \left(\frac{H(Y) - H(Y \mid X)}{H(Y) + H(X)} \right)$$

We will refer to symmetric uncertainty between attribute X and UXO as *SU*(X;UXO).

Next, Maximum Relevance Minimum Redundancy methods ("MRMR") are used to locate attribute sets with the maximum amount of information between the attributes and UXO classification and simultaneously, the minimum amount of overlapping information as between the individual attributes in the dataset.[23] In other words, MRMR does not look for just the best attributes measured by information between the individual attributes and UXO classification. Such attributes are frequently highly correlated and contain very similar information about the target output. Having five such attributes adds to an existing attribute contributes little or nothing to solving the problem. Rather, MRMR attempts to construct the attribute set that collectively contains the most information about the dependent variable.

MRMR is a greedy best-first algorithm. That is, it searches the entire attribute set for the single attribute that best increases the Relevance/Redundancy objective function. That attribute is added to the attribute set and that decision is not reexamined. Then the MRMR algorithm searches for the next attribute that, when added to the existing selected attribute set best maximizes the objective function.

---

[23] Hanchuan Peng, Fuhui Long, and Chris Ding (2005). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp.1226-1238.

6.9.3.2 Attribute Reduction Process

Using the MRMR, we were able to reduce the data set to a 19 attribute set. Then, the pseudo principal components (sums and differences of two attributes) were derived for each pair-wise combination of the 19 reduced attributes, creating additional attributes that needed to be reduced.

These attributes were reduced again using a J48 single decision tree algorithm which is an extension of the classic C4.5 decision tree algorithm.[24] J48 builds decision trees from a set of labeled training data using information entropy. Each attribute of the data can be used to make a decision by splitting the data into smaller subsets. The J48 algorithm may be summarized as follows:

> "J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class."[25]

J48 was used as an alternative way to pick out attribute sets from MRMR and CFS. J48 is stronger at picking out interactions amongst attributes than is either MRMR or CFS. The following J48 tree was created (we note that items like "BJ", "FQ" and the like are designators of attributes in the 19 member attribute set).

J48 Pruned Tree:

BJ <= -0.106768

| FQ <= -0.828438: 0 (62.0)

| FQ > -0.828438

| | HM <= -0.048875

| | | M <= 0.46244: 1 (4.0)

| | | M > 0.46244: 0 (3.0)

| | HM > -0.048875: 0 (19.0)

BJ > -0.106768

| CR <= -2.638582: 0 (13.0)

| CR > -2.638582

| | L <= 1.065826

| | | GH <= 0.627855: 0 (2.0)

| | | GH > 0.627855: 1 (29.0/1.0)

---

[24] Quinlan, Ross (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[25] http://www.opentox.org/dev/documentation/components/j48

| | L > 1.065826: 0 (6.0/1.0)

The attributes from the above tree (BJ, FQ, HM, M, CR, L, and GH) were then used as inputs to a Random Forests[26] modeling algorithm. Random Forests is an ensemble decision tree algorithm that is reasonably fast and is does a generally good job of building reasonably robust models. We use Random Forests to assess the probable predictive result of a particular attribute set and also use its variable importance rankings as an attribute excluder. We performed 50-fold cross-validation, each fold comprising a forest of 1000 trees and compared the results across different attribute combinations. The most favorable Random Forests model of that group resulted from using attributes BJ, FQ and HM. In selecting these attributes, we consciously favored models that found the final UXO earlier over models with a higher area under the curve. The forgoing three attributes are the attributes that were used for LGP modeling on the first iteration.

The cross-validation results from the Random Forests models were:

Random Forest:

Test mode: 50-fold cross-validation

Random forest of 1000 trees, each constructed while considering 2 random features.

Out of bag error: 0.1594

Summary:

Correctly Classified Instances	113	81.8841 %
Incorrectly Classified Instances	25	18.1159 %
Kappa statistic	0.4631	
Mean absolute error	0.1991	
Root mean squared error	0.3248	
Relative absolute error	54.32%	
Root relative squared error	75.96 %	
Total Number of Instances	138	

Detailed Accuracy By Class:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.914	0.485	0.857	0.914	0.885	0.906	0
0.515	0.086	0.654	0.515	0.576	0.906	1
0.819	0.389	0.809	0.819	0.811	0.906	Weighted Avg.

Confusion Matrix:

		Predicted	
		Not-UXO	UXO
Actual	Not-UXO	96	9
	UXO	16	17

[26] Random Forests™ is a trademark of Leo Breiman.

Accordingly, BJ, FQ and HM were used for further modeling. They may be described as follows:

1. Attribute BJ = Second pseudo principal component of the:
    - Base 10 log of the second moment in the inner part of the ellipse for the final decay channel; and
    - First moment of the ratios of the third decay channel and fourth decay channel in the inner part of the ellipse.
2. Attribute FQ = Second pseudo principal component of the:
    - Third moment of the first decay channel in the ellipse; and
    - Second moment of the ratios of the second decay channel to the third decay channel in the inner part of the ellipse.
3. Attribute HM = Second pseudo principal component of the:
    - Third moment of the ratios of the third and fourth decay channels in the ellipse; and
    - Third moment of the second decay channel in the inner part of the ellipse.

6.9.3.3 Selected Attribute Match between Training and Blind Data

Since the training data will be used to make predictions to the blind data, we looked at how well the training and blind data "match". We did this by looking at the descriptive statistics and graphs for all three attributes individually. The descriptive statistics are shown in Table 16, Table 17, and Table 18 and the graphics in Figure 30, Figure 31, and Figure 32. It was determined that training and blind data were reasonably consistent.

Table 16. Descriptive statistics for attribute BJ

Attribute BJ											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1081	-0.19646	0.42675	0.18212	0.5573	-0.29102	2.65717	0.628	0.000*	-0.012	>.10
Training Data	138	-0.16875	0.45254	0.2048	0.64484	-0.28899	2.07306	0.53	0.012*	-0.445	>.10
Blind Data	943	-0.20052	0.42295	0.17888	0.53733	-0.29433	2.65717	0.643	0.000*	0.068	>.10

Figure 30. Box and whisker plots for attribute BJ

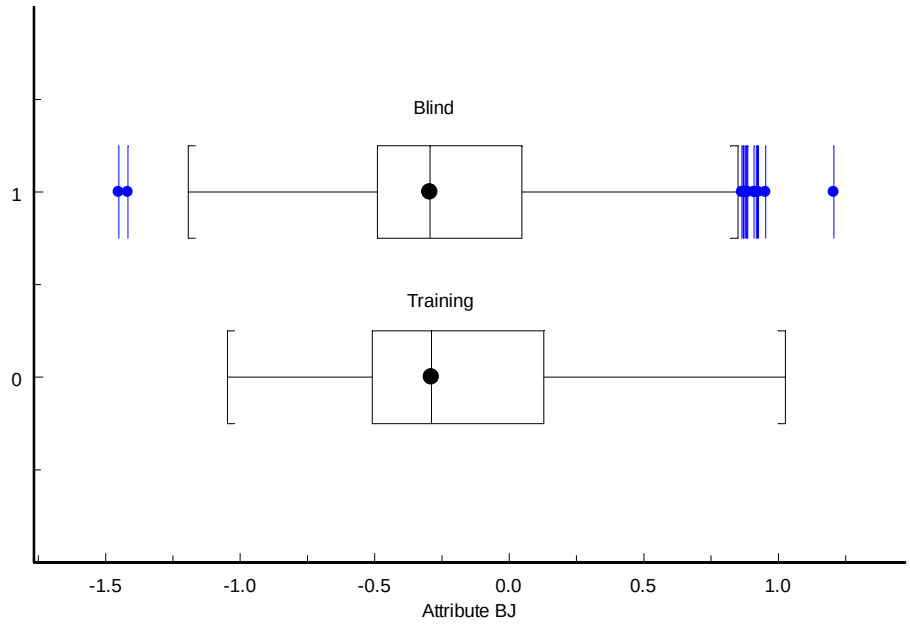


Table 17. Descriptive statistics for attribute FQ

Attribute FQ											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1081	-0.8858	0.15297	0.0234	0.20755	-0.87449	1.46532	-0.343	0.000*	1.091	<.02*
Training Data	138	-0.88043	0.14665	0.02151	0.1827	-0.87232	1.02131	-0.674	0.002*	2.079	<.02*
Blind Data	943	-0.88658	0.15394	0.0237	0.20975	-0.87486	1.46532	-0.3	0.000*	0.988	<.02*

Figure 31. Box and whisker plots for attribute FQ

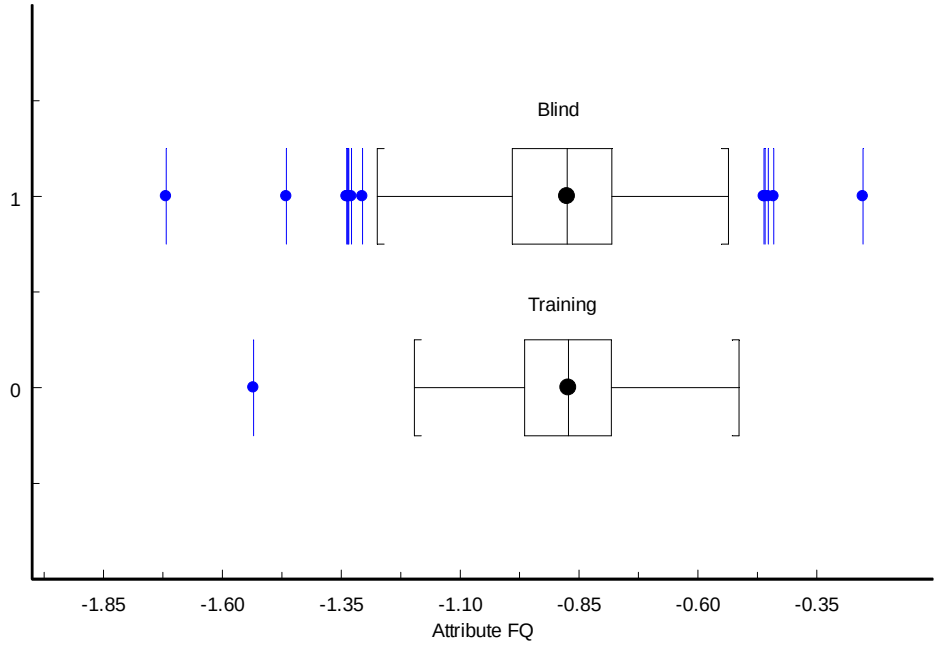
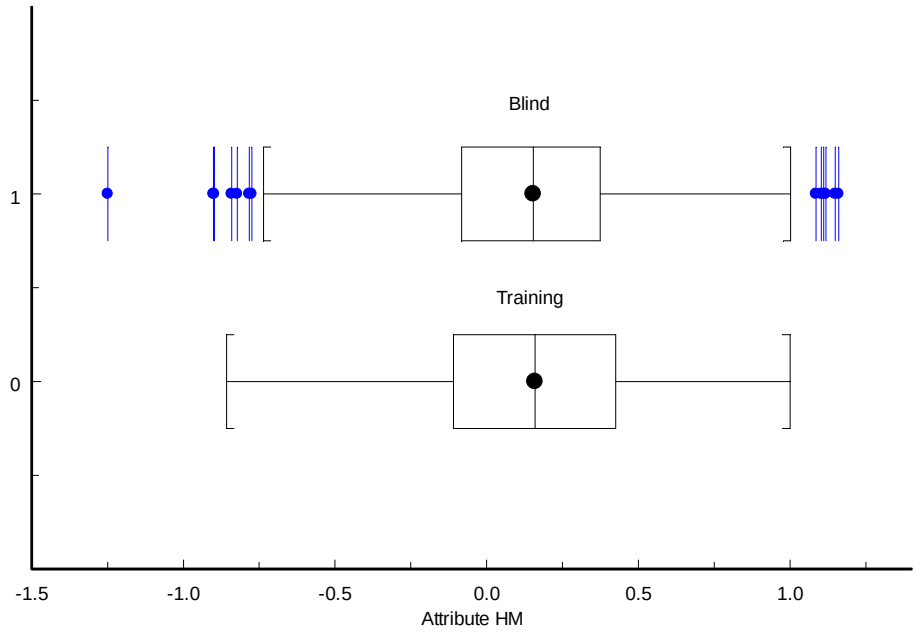


Table 18. Descriptive statistics for attribute HM

Attribute HM											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1081	0.13984	0.3474	0.12069	0.45904	0.15507	2.40988	-0.09	0.226	0.185	>.10
Training Data	138	0.14731	0.36385	0.13239	0.53902	0.15973	1.85725	-0.048	0.813	-0.117	>.10
Blind Data	943	0.13875	0.34511	0.1191	0.45673	0.15413	2.40988	-0.098	0.217	0.241	>.10

Figure 32. Box and whisker plots for attribute HM



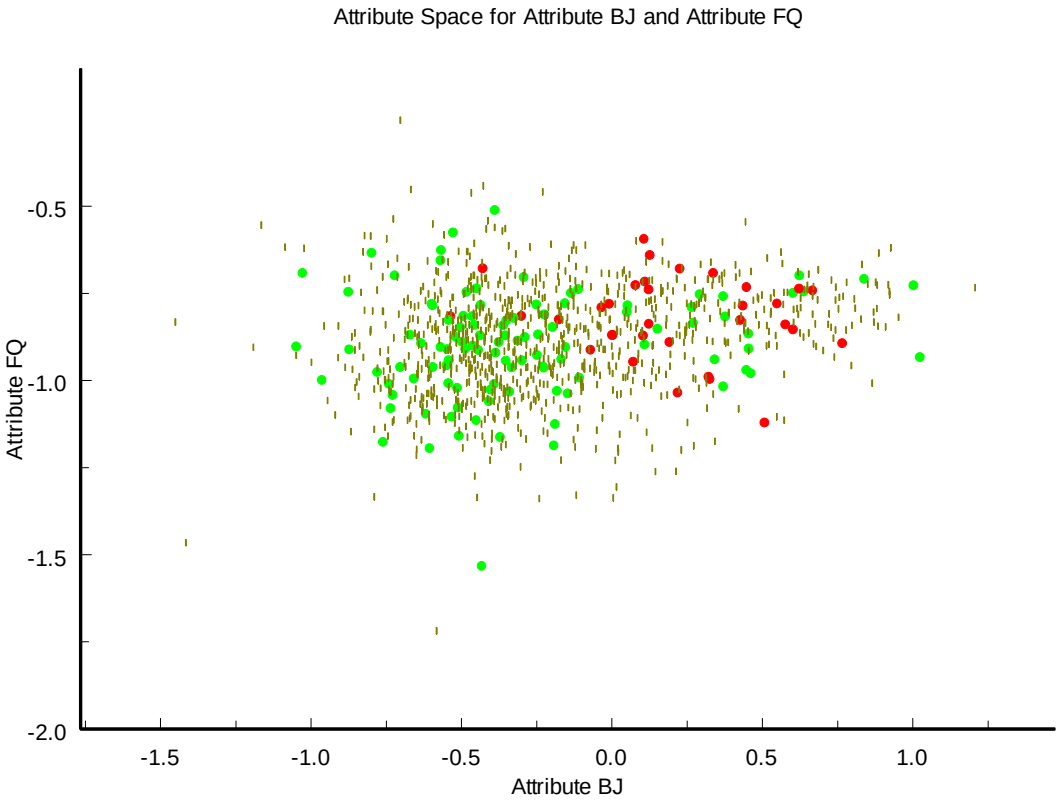
6.9.3.4 Selected Attribute Space Graphs

The graphs in this section show three different views of attribute space, given the three selected attributes, BJ, FQ and HM. In these graphs, red circles represent UXO in the training data, green circles represent Not-UXO in the training data and brown lines represent the blind data. Figure 33, Figure 34, and Figure 35 show: (1) A reasonably good match between the training and blind data in multidimensional space; and (2) Good separation of UXO and Not-UXO.

6.9.3.4.1 Attribute BJ versus Attribute FQ

Figure 33 shows a plot of attribute space for attributes BJ and FQ.

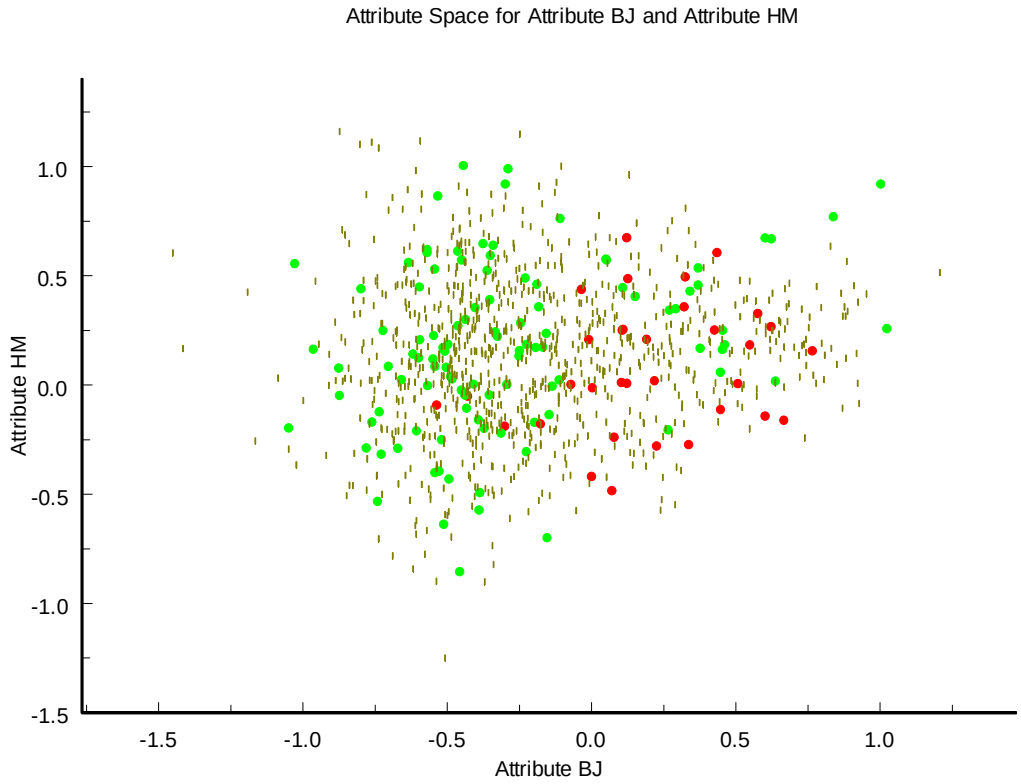
Figure 33. Attribute space for attribute BJ versus attribute FQ (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind targets).



6.9.3.4.2 Attribute BJ versus Attribute HM

Figure 34 shows attribute space for attributes BJ and HM.

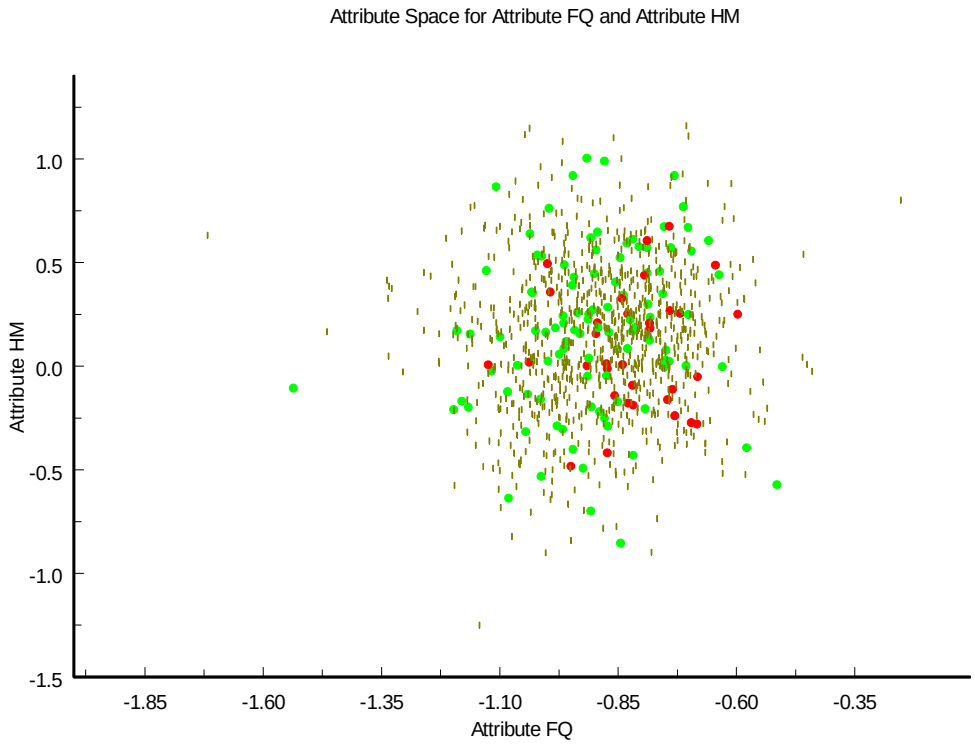
Figure 34. Attribute space for attribute BJ versus attribute HM. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind Data).



6.9.3.4.3 Attribute FQ versus Attribute HM

Figure 35 shows attribute space for attributes FQ and HM.

Figure 35. Attribute space for attribute FQ versus attribute HM. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind Data).



6.9.4 LGP Discriminator

The LGP Discriminator was used to develop a model that predicts the ranking of blind targets in order of the likelihood they are UXO or Not-UXO.

6.9.4.1 Training Set Used

For training the model we used all training targets not assigned to cannot-analyze one, two or three and not assessed as high-probability Not-UXO by the amplitude discriminator.

6.9.4.2 Parameters used for Deriving LGP Model on the Training Data

Discipulus™ 5.0 was used to evolve our LGP model on the remaining training data using the three Attributes (BJ, FQ and HM) as our input variables (Independent Variables) and UXO classification as our output variable (Dependent Variable).

This is a small data set. The biggest danger is overfitting to the training data and producing models that do not generalize well to the blind data. Attribute reduction was the first important step to preventing overfitting. Here are the additional steps to parameterize our LGP projects models and minimize the danger of overfitting.

The most important modeling decision was that the data set was small enough that we should add noise to the attributes to prevent overfitting. We replicated each row in the training data 30 times and added a small amount of noise to each input, defined by a percentage –from 1% to 10%.

Adding noise in inductive modeling is equivalent to Tikhonov Regularization and, if the correct noise level is selected, reduces overfitting.[27]

We selected the noise parameter by performing twenty-fold cross-validation projects at selected noise levels from 3% to 10% in 1% increments. Discipulus was set to its default parameters, except for the following: (1) Fitness function was set to "Ranking-Best ROC Curve"; (2) Each run in the project was terminated at 150 generations without improvement; and (3) The number of runs in each project was 20. At the end of each project/fold, we opened the program designated by Discipulus as the best program of the project and we repeatedly removed introns from that program until the best program ceased getting shorter. The best program with introns removed was selected as the program model for that fold. Its scores on the held-out data for that fold were stored. After all twenty cross validation projects were completed, the stored scores were aggregated and targets with multiple scores were assigned a score equal to the average score for that target. This provided a single score for every target, which we interpret as a ranking.

Table 19 shows the twenty-fold cross-validated results of LGP projects in terms the area under the curve (higher is better) and the count of misranked Not-UXO--that is, how many Not-UXO were ranked above the lowest ranked UXO by noise level. Obviously, a lower value is better on the count of misranked.

Table 19. Performance of various noise levels in iteration one using twenty-fold cross-validation

Iteration 1		
Noise	**AUC**	**Misranked**
3%	86.71%	89
4%	90.44%	36
5%	86.28%	58
6%	87.03%	72
7%	86.63%	34
8%	88.84%	39
9%	85.29%	75
10%	87.12%	73

Based on these results, we selected 4% and 7% as the noise levels with which to perform further modeling on this iteration.

Other parameter settings used in LGP modeling projects were:

The fitness function used for all LGP modeling runs was "Ranking-Best ROC Curve" as measured by area underneath the curve.

The parameter settings for all LGP modeling runs were the default Discipulus™ 5.0 parameters with the following changes:

Stepping = Disabled

[27] Bishop, C. (1995) "Training with Noise is Equivalent to Tikhonov Regularization." *Neural Computation* 7 No. 1 (1995) 108-116.

Single Run Termination:

Generations without Improvement = 150

Use Adaptive Termination = Disabled

Batch Run Termination: Maximum Number of Runs = 20

Single Run Parameters:

Population Size = No Randomization; Set to 500

Maximum Program Size = No Randomization; Set to 128

Subset Size = No Randomization

6.9.4.3 Creating the LGP Ensemble Predictor

We then performed fifty bagging projects with Discipulus LGP using each of the selected noise parameters with in-bag set size equal to the size of the training input set. For Discipulus, we used the same parameters described above for the cross-validation runs.

At the end of each project/bag, we opened the program designated by Discipulus as the best program of the project, we repeatedly removed introns from that program until the best program ceased getting shorter. The best program with introns removed was selected as the program model for that bag. Its scores on the out-of-bag data for that fold were stored. After all one-hundred bagging projects were completed, the stored scores were aggregated and targets with multiple scores were assigned a score equal to the average score for that target. This provided a single score for every target, which we interpret as a ranking.

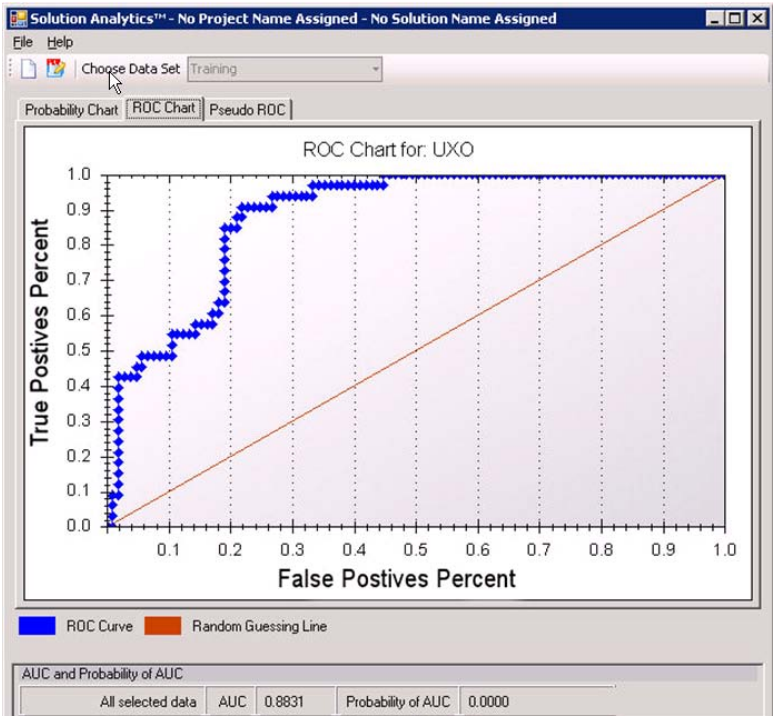
In addition, after each project/bag was completed, we stored the scores of the same program on the blind targets. This produces multiple scores for each target. The average score for each blind target was treated as the predictive ranking for that target.

At the end of this process, we had constructed an "LGP ensemble predictor," comprised of one-hundred evolved programs from LGP, each of which had been trained on a different sample from the training data set. The outputs from those one hundred programs was reduced to a single predictor for the training and blind targets.

6.9.4.4 Performance of LGP Ensemble Predictor on Training Data

The area under the curve of the ROC curve on the training data of the derived model was 88.31%. These results are shown only on the out-of-bag training data summed across all bags. Out-of-bag data is not used to train the model. Figure 36 shows the ROC curve generated by the derived LGP model on the out-of-bag training data.

Figure 36. ROC curve on training data for LGP model (shown without cannot-analyze targets)



6.9.4.5 Determination of Cannot-Analyze Four Category Targets

In looking at the attribute space graphs we decided to send five additional blind targets to a "cannot-analyze four" category of targets because they appeared to be blind targets that are outliers in attribute space. Figure 37 shows two dimensions of outlier space and shows four of the removed targets. The removed outliers are highlighted. The removed targets were obviously selected because they were isolated blind data points for which there was no nearby training data and where, even if we sampled them after iteration one, they would yield no useful information about nearby blind targets.

Figure 37. BJ vs FQ attribute space with outliers designated

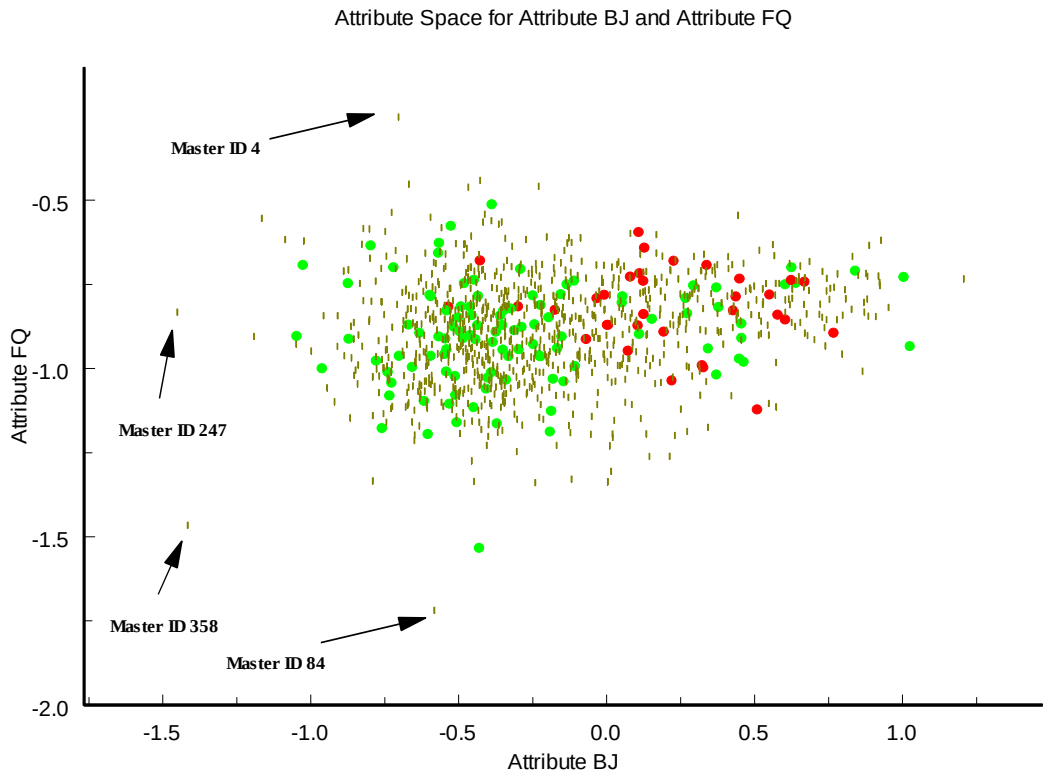


Table 20 shows the updated sample sizes after we removed the five cannot-analyze four targets from the blind dig list.

Table 20. Count of targets lost due to cannot-analyze four (attribute space outliers)

Iteration 1	Train	Blind	Total
Original Data Sample Size	182	1282	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	157	1072	1229
Cannot Analyze 2 (Low Sample Size)	(1)	(1)	(2)
Sample Size After Cannot-Analyze 2 Removed	156	1071	1227
Amplitude Discriminator Low Prob UXO	(17)	(118)	(135)
Sample Size After Amplitude Discriminator Removed	139	953	1092
Cannot Analyze 3 (Low Sample Size -Round 2)	(1)	(10)	(11)
Sample Size After Cannot-Analyze 3 Removed	138	943	1081
Cannot Analyze 4 (Outliers)	0	(5)	(5)
Sample Size After Cannot-Analyze 4 Removed	138	938	1076

6.9.5 Risk Analysis

The LGP ensemble predictor generates a ranking for all blind targets. The next step in this process is to determine the LGP ensemble predictor stop-digging threshold. The LGP values were first converted into ranks across the entire training and blind data. Lower LGP values were

interpreted as larger rankings. Next, kernel regression with a Gaussian kernel was used to model the Probability of UXO for each target using the training data. The functional form for kernel regression is:

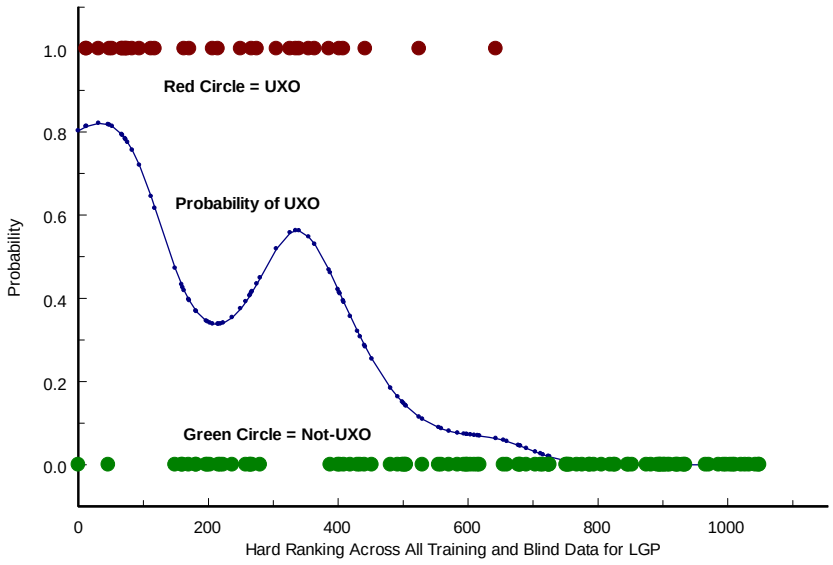
$$\mathrm{P}(UXO)_i = \sum_j e^{-\left(\frac{(x_i - x_j)^2}{2\alpha^2}\right)}$$

Where: (1) α represents the standard deviation of the Gaussian kernel (the single settable parameter); (2) x_i represents rank of the ith ranked blind data instance computed from the LGP scores across all training and blind data points; and (3) x_j represents rank of the jth ranked training data instance value of the LGP values across all training and blind data points.

The value determined for the parameter, α, is 59.37. That value was determined by n-fold cross-validation on the training data. The α parameter selected was one that produced the minimal value for -2*log likelihood over the training data, which is the maximum likelihood estimator for these data, assuming Bernoulli errors.

Figure 38 shows the derived kernel regression model—probability of UXO is plotted as a function of rank on the training data. Note that the rankings are derived from the LGP values and represent the rankings across all training and blind data not assigned to cannot-analyze one, 2, 3 or that were removed by the amplitude discriminator.

Figure 38. Kernel regression fit between probability of UXO and LGP rank on training Data



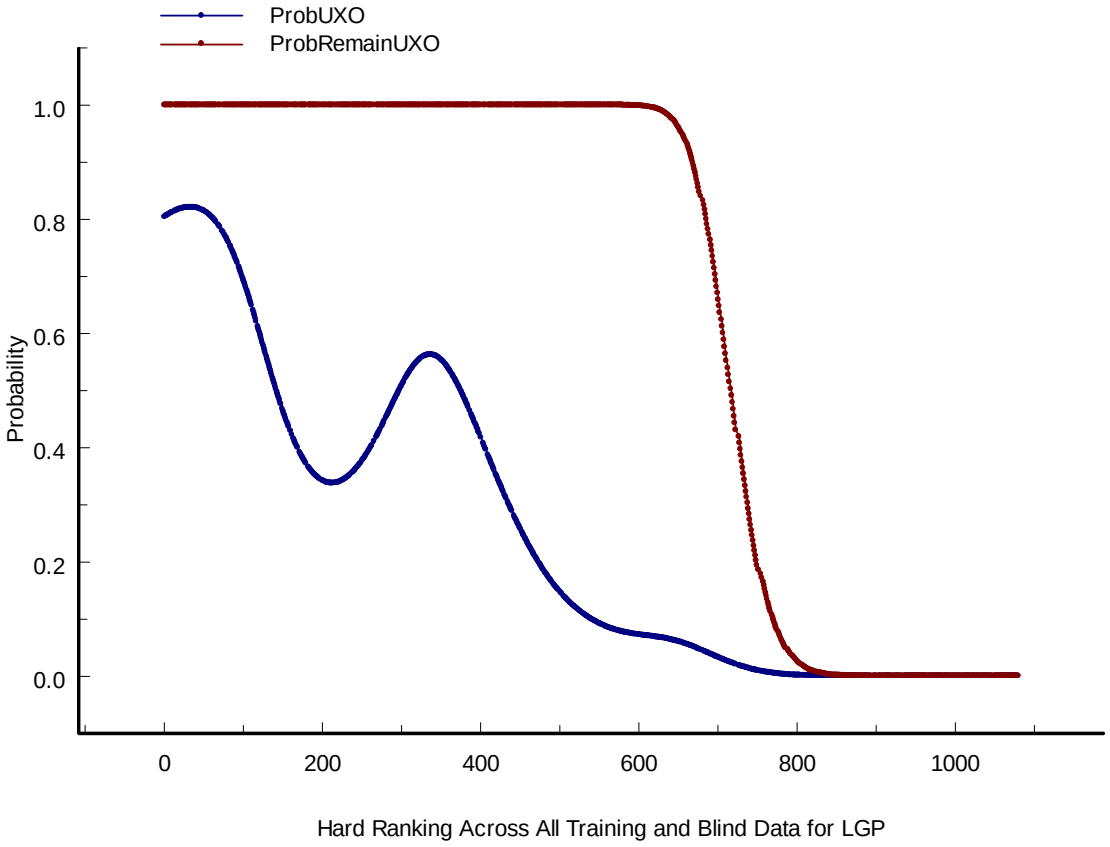
The Gaussian kernel, generated by the training data, using the above kernel width parameter, was then applied to the ranked blind data, generating a probability that each blind data item is UXO.

Once individual target probabilities are set, the probability that all blind targets above each LGP ranking contain one or more UXO is calculated using the approach outlined in 2.1.6. This is the residual risk as a function of rank. In particular, we used Equation 1 and Equation 2 to compute

the OR of the probabilities for all targets from the ranking for which the computation is being performed to the most extremely ranked blind target.

The blue line in Figure 39 is the probability of UXO as a function of the rank across all training and blind predictions derived from the LGP values. The red line is the probability that one or more UXO remain on site at each rank value in the blind targets, assuming all items to the left of that rank have been excavated. When the red line reaches a critical probability value ($p\text{-value}_{crit}$), we assess all targets remaining to the right of that rank (i.e. targets with a larger rank) as safe to leave in the ground.

Figure 39. Kernel regression applied to blind data



A 95% confidence level was chosen at the start of this project and is used throughout this project. Since there were two risk analyses used to determine two stop-digging thresholds the Bonferonni correction needed to be applied.[28] Using the Bonferonni correction the p-value$_{crit}$ was set to .025 (2.5%).

The p-value$_{crit}$ was then used to determine the critical rank value of 802, inferring that any target with a rank greater than 802 was high-probability Not-UXO. At that point, the probability of remaining UXO was 0.02437—in other words, it satisfies the p-value$_{crit}$ criterion.

Table 21 shows the details for the 95% stop-digging threshold.

Table 21. Stop-digging threshold

LGP						
Cut-Off	**TID**	**Score**	**HardRank**	**ProbUXO**	**ProbRemainUXO**	**% Of Blind Data Left in Ground**
95.00%	519	0.279664552	802	0.00138	0.02437	18.49%

The blind data targets that fell below the amplitude discriminator stop-digging threshold were combined with the blind data targets that fell below LGP discriminator stop-digging threshold. This below threshold list contains the blind data targets identified as safe to leave in the ground. We then calculated the total % of blind data left in the ground by summing the percentage left in the ground using the amplitude discriminator (9.20%) and the percentage left in the ground using the LGP discriminator (18.49%). This results in 27.69% targets out of the original 1282 blind targets that can be left in the ground for iteration one.

6.9.6 Prepare Prioritized Dig List

To assemble our prioritized dig-list we had to assemble the targets assessed as high probability Not-UXO by the amplitude discriminator and with all of the targets scored by the LGP ensemble predictor. We assembled below dig threshold targets together and ranked them by the probability generated for the target by the risk analysis model that assigned them to "do-not-dig." The above stop-digging threshold targets were ordered by the probability of remaining UXO assigned to the targets by the risk analysis model that used the LGP ensemble predictor scores for ranking.

When complete, the dig-list provided a ranking, Master ID, and a label whether the target was above or below the stop-digging threshold or, alternatively, a cannot-analyze target.

6.10 Request for Further Ground-Truth

After iteration one was complete, we elected to select more ground-truth to improve the training set and improve discrimination. This would be the equivalent, on an actual site cleanup, of requesting that additional targets be dug and including those targets in additional models and risk analysis.

The number of examples of UXO in the iteration one training set was very small and we expected improvement from a larger sample size. In the experimental plan, we were limited to 20% of the blind data for sampling. Given these data, that permitted us to select 256 targets. Because of the time constraints of this project, we did not believe more than two iterations were possible. Accordingly, we determined to sample all additional ground-truth (the 256 targets) for

[28] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

iteration two. We took our samples after removal of the Cannot Analyze One targets were removed. This is reflected in Table 22

Table 22. Iteration 2 original data sample size

Iteration 2	Train	Blind	Total
Original Data Sample Size	438	1026	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	413	816	1229

Sampling additional ground-truth was performed by four criteria. Those criteria are listed below along with the number of samples allocated to each criterion. The targets selected were forwarded to the Program Office and they returned the ground-truth for them to us.

The four sampling criteria were:

6.10.1 Entropy

Entropy is a measure of the uncertainty of an unknown target. The more uncertain we are about a blind target, the more information we get from sampling that target. We selected 95 blind targets by this criterion. To do so, we selected 95 targets randomly with the random selection weighted by entropy. The higher the entropy of a target, the more likely it was to be selected. Entropy was computed from the probability each target is UXO generated by our risk analysis from iteration one.

The expected cost of digging a target is just one minus the probability that target is UXO. This measures the probability that we are digging something that could be safely left in the ground. The expected cost of these 95 samples is 49.9 Not-UXO dug.

6.10.2 Entropy per Unit of Expected Cost of Sample

Entropy per Unit of Expected Cost is a criterion designed to get looks at likely UXO at the lowest possible cost. In other words, entropy measures expected information content and expected cost measures the likelihood that we are digging Not-UXO. Thus entropy per unit of expected cost looks for the targets that provide “cheap” information. We use expected cost and entropy as described above and sampled 60 targets by this criterion, again, using a weighted random sample.

The expected cost of these 60 targets is 21.6 Not-UXO dug.

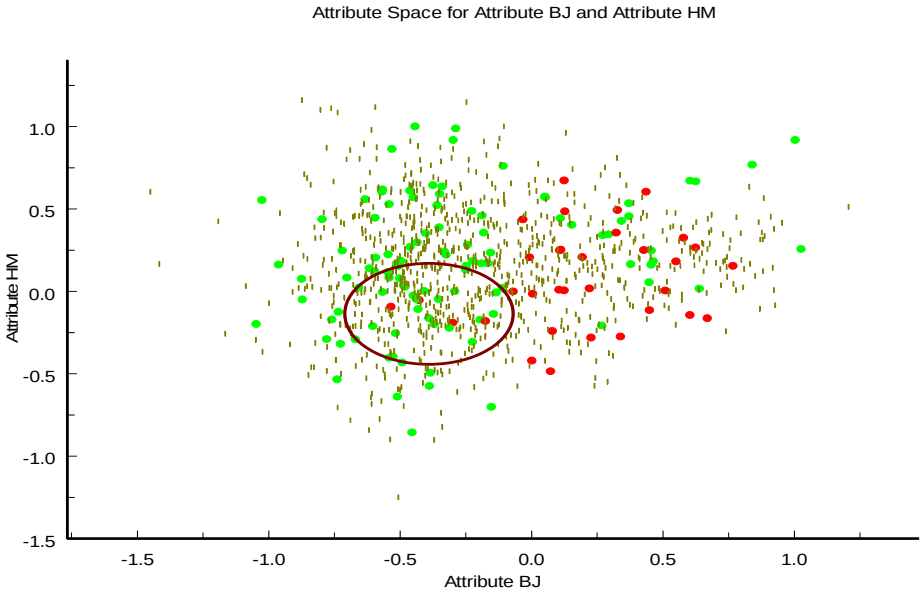
6.10.3 Visual Picks around Training Outliers

During the iteration one training, three training UXO targets consistently stood out as more difficult to discriminate than the remainder, 22, 36, and 410. We picked blind targets manually in the immediate vicinity of these targets in attribute space to sample. The goal was to determine if these were just outliers or actually represented a cluster of UXO that was poorly represented in the training data. Altogether, we picked 50 samples by this criterion.

Figure 40 shows the three outliers in red. The brown ellipse designates the region around those outliers from which we sampled.

The expected cost of these samples was 38.5 Not-UXO dug.

Figure 40. Region of selection of blind targets for sampling around an outlier UXO



6.10.4 Random Sample from Tail of Risk Analysis Probability

The rankings on our dig list between the last training UXO and the dig threshold comprise a region in which we wish to acquire more information so that the tail of the declining probability is better defined. We sampled 50 blind targets randomly out of this region.

The expected cost of these 50 targets is 47.3 Not-UXO dug.

6.10.5 Expected vs. Actual Cost

The total expected cost of our request for ground-truth was 157.3 Not-UXO and the expected number of UXO was 98.7. When we received the ground-truth, the actual cost was 162 Not-UXO and the number of UXO found was 94.

6.11 ITERATION TWO

6.11.1 Introduction

Iteration two proceeded immediately following the receipt of our request for more ground-truth. We did not repeat our preliminary steps described above such as extraction of polygons, definition of ellipses, definition of cannot-analyze one targets, attribute extraction or definition of cannot-analyze two category targets. Iteration two started from the same point as iteration one. See Sections: 6.3, 6.4, 6.5, 6.6, 6.7, and 6.8. The only difference was more training ground-truth.

Further, iteration two proceeded in a series of steps almost identical to iteration one, except in implementation details. Those steps and implementation details are described below.

6.11.2 Description of Data

The DGM from the SLO EM61MTADS data, processed as described above for iteration one was used for Iteration two.

The training ground-truth used in iteration two is set forth in Table 23:

Table 23. Ground-truth labels for SLO iteration two

Iteration 2		
Type	**Count**	**%**
Not-UXO	316	72.15%
60mm Mortars	47	10.73%
80mm Mortars	24	5.48%
2.36" Rockets	11	2.51%
4.2" Mortars	24	5.48%
5" Rocket Warhead	1	0.23%
RML UXO	15	3.42%
Total	438	

After removal of the cannot-analyze one and two category targets, Table 24 shows the training, blind and total sample sizes we worked with in this iteration.

Table 24. Training and blind data counts for iteration two

Iteration 2	Train	Blind	Total
Original Data Sample Size	438	1026	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	413	816	1229
Cannot Analyze 2 (Outlier)	(1)	0	(1)
Sample Size After Cannot-Analyze 2 Removed	412	816	1228

6.11.3 Amplitude Discriminator

6.11.3.1 Introduction

We had found it useful in previous work to filter targets with an amplitude discriminator. An amplitude discriminator is a single attribute derived using Amplitude Statistics that, by itself, discriminates. That is, the derived attribute itself is the model.

The purpose of this was to use a very simple initial model to identify a set of targets that may be excluded as high-probability Not-UXO. The purpose of using only Amplitude Statistics is our observation that they are more robust against noise than the Ratio Statistics. Once excluded, the distributions of the Ratio Statistics stabilize and provide more robust prediction

6.11.3.2 Attribute Identification

The same process used in iteration one was used here, that is, all of the extracted attributes were binned using a Chi-square optimal binning procedure. Once binned, the count of UXO in the low and high bins were examined for each attribute to determine which had the lowest number of UXO in the low (or high) bin and the highest number of Not-UXO in the low (or high) bin.

Using this procedure two attributes were identified as possible candidates for the amplitude discriminator. We combined attributes 1 and 2 using principal components analysis and used the first principal component, Attribute AD2 as the iteration two amplitude discriminator.

Attribute AD2 may be described as the first principal component between:

- Attribute 1: Base 10 log of the first moment of the fourth decay channel in the ellipse; and

- Attribute 2: Third moment of the first decay channel in the ellipse.

Table 25. Bin Information for Attribute 1

Attribute 1					
Bin	**Bin Boundary**	**Bin Boundary**	**Total Targets**	**UXO Targets**	**Non-UXO Targets**
Bin 1	-INF	54.28999996	88	0	88
Bin 2	54.28999996	55.78000021	85	12	73
Bin 3	55.78000021	58.77999973	114	32	82
Bin 4	58.77999973	INF	124	77	47

Table 26. Bin Information for Attribute 2

Attribute 2					
Bin	**Bin Boundary**	**Bin Boundary**	**Total Targets**	**UXO Targets**	**Non-UXO Targets**
Bin 1	-INF	0.250342034	88	2	86
Bin 2	0.250342034	0.444245819	127	27	100
Bin 3	0.444245819	INF	196	92	104

6.11.3.3 Attribute Match between Training and Blind Data

Since the training data will be used to make predictions to the blind data, we reviewed the match between the distributions of the training and blind data on the selected attribute. We reviewed the descriptive statistics in Table 27, Table 28, and Table 29; and the charts in Figure 41, Figure 42, and Figure 43.

Those tables and figures demonstrate that the training data had a larger sample mean and variance then blind data for Attribute 1, Attribute 2, and attribute AD2.

This is almost certainly due to the fact that our new sampled ground-truth was not a 100% random sample. Fifty of the 256 new ground-truth targets were chosen randomly and the other 206 targets were chosen purposely to maximize the information yield of the sample. This meant the sample was biased, by its nature, toward proportionally more UXO than the site as a whole. Accordingly, we expected a tendency for the training data to be distributed to look more like UXO than the blind data. That is precisely what the following tables and charts show. We do not, however, see extreme blind outliers in regions not represented by training data and accordingly, we accepted attribute AD2 as the amplitude discriminator attribute.

Table 27. Descriptive Statistics for Attribute 1 (Used in AD2)

Attribute 1 (Used in AD2)											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1228	1.75356	0.03264	0.00107	0.03526	1.74556	0.23008	1.561	0.000*	3.407	<.02*
Training Data	412	1.75938	0.03358	0.00113	0.03866	1.75168	0.22766	1.25	0.000*	1.978	<.02*
Blind Data	816	1.75062	0.03178	0.00101	0.03252	1.74241	0.22786	1.774	0.000*	4.594	<.02*

Figure 41. Box and Whisker Plots for Attribute 1 (Used in AD2)

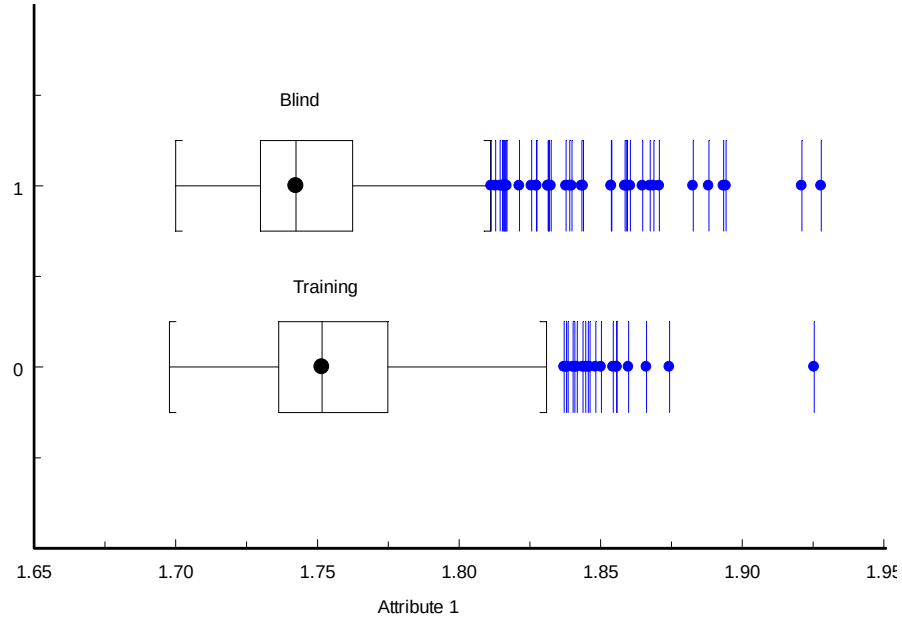


Table 28. Descriptive Statistics for Attribute 2 (Used in AD2)

Attribute 2 (Used in AD2)											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1228	0.3491	0.23734	0.05633	0.33915	0.35988	1.45185	-0.411	0.000*	-0.098	>.10
Training Data	412	0.41175	0.2125	0.04516	0.30121	0.43237	1.45185	-0.562	0.000*	0.741	.10-.05
Blind Data	816	0.31747	0.243	0.05905	0.34517	0.32819	1.33835	-0.302	0.001*	-0.338	>.10

Figure 42. Box and Whisker Plots for Attribute 2 (Used in AD2)

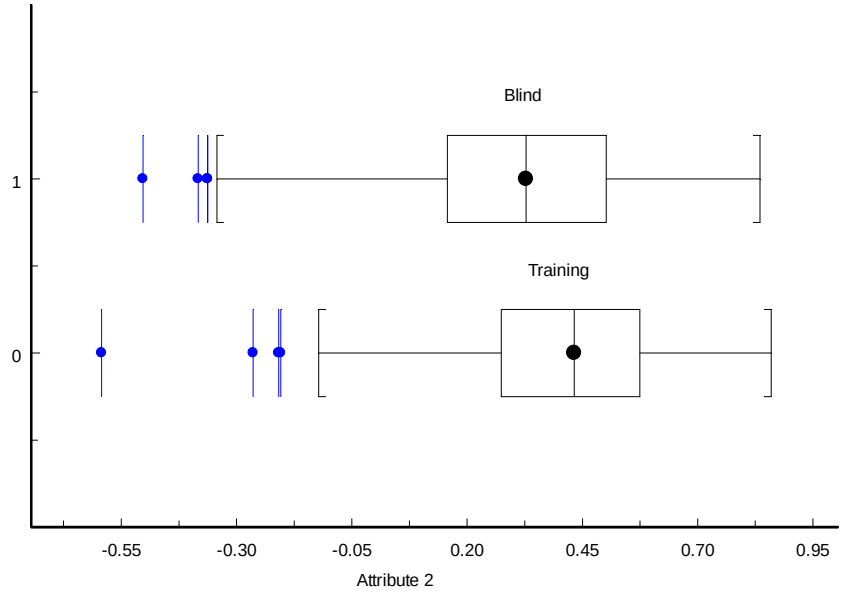


Table 29. Descriptive Statistics for Attribute AD2

Attribute AD2											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	1228	0	1.15409	1.33191	1.40136	-0.15736	8.69584	0.651	0.000*	0.642	<.02*
Training Data	412	0.31285	1.14032	1.30033	1.53762	0.10819	8.69584	0.417	0.001*	0.51	>.10
Blind Data	816	-0.15796	1.12917	1.27502	1.32059	-0.30889	7.5345	0.815	0.000*	1.017	<.02*

Figure 43. Box and Whisker Plots for Attribute AD2

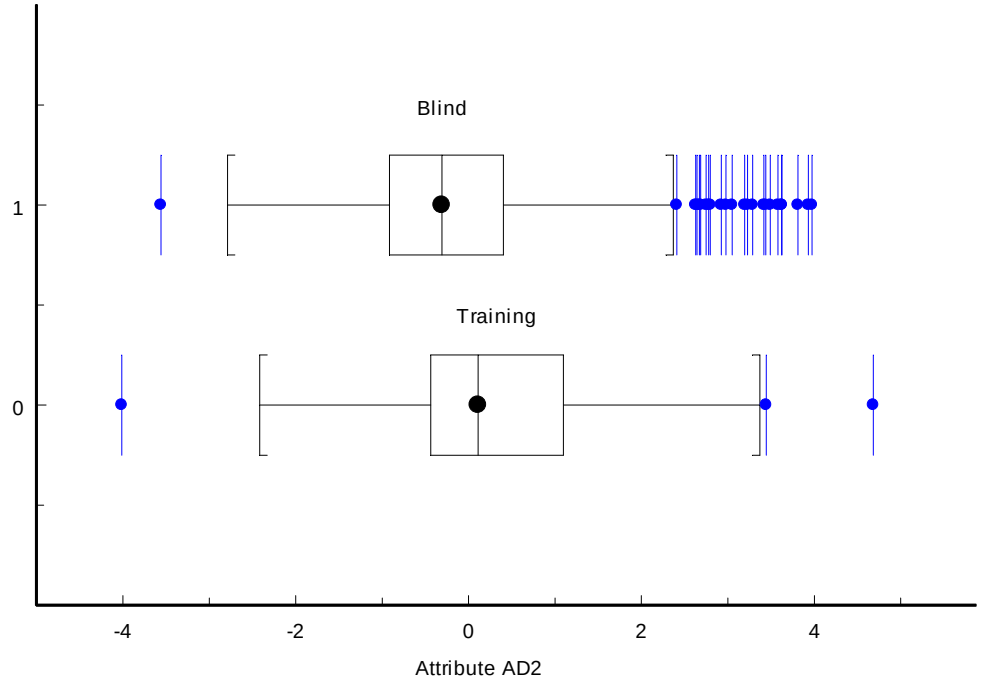
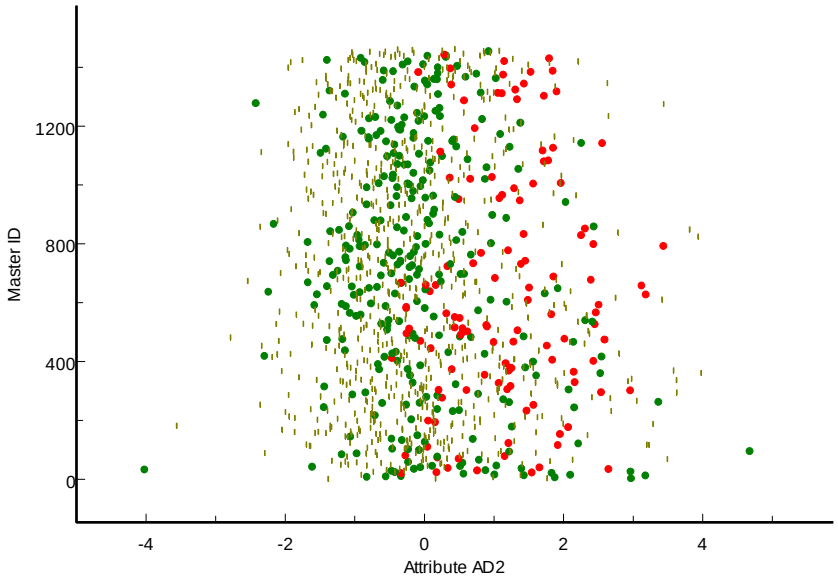


Figure 44 shows how UXO and Not-UXO are distributed in the one dimensional attribute space represented by Attribute AD2. Note that the y-axis in this figure is just the Master ID. It is used solely to spread the data points out for easier visualization. The AD2 attribute as a discriminator is simple to understand. The larger the value of Attribute AD2, the more likely an item is to be UXO. Thus, AD2 provides a ranking using a single dimension.

Figure 44. Attribute Space for Attribute AD2 (Red circles are UXO. Green circles are Not-UXO. Brown Lines are Blind Data. Y-axis is Master ID. It is used solely to spread out the values for better visualization).



6.11.3.4 Risk-Analysis/Stop Digging Threshold

The next step in this process is to determine the stop-digging threshold, given that we are using AD2 as a ranker. The AD2 values were first converted into ranks across the entire training and blind data. Lower AD2 values were interpreted as larger rankings. Next, kernel regression with a Gaussian kernel was used to determine the Probability of UXO for each target:

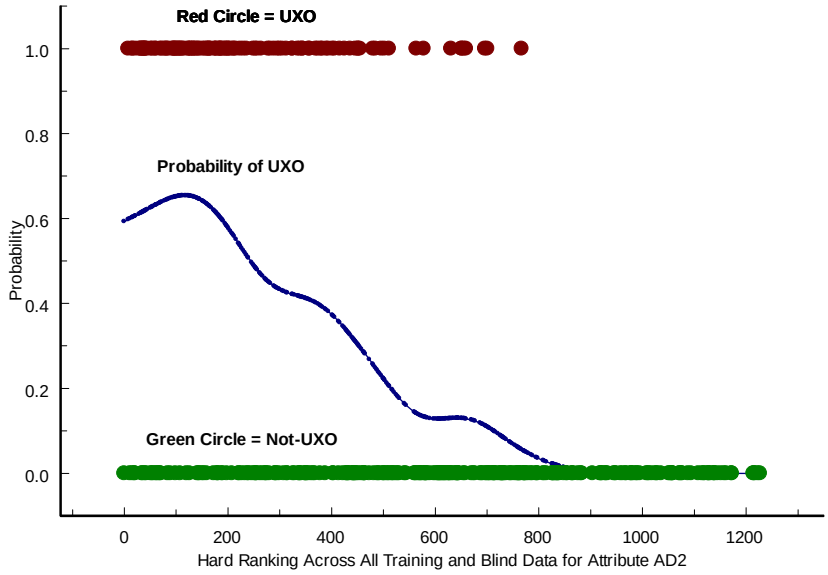
$$\mathrm{P}(UXO)_i = \sum_j e^{-\left(\frac{(x_i - x_j)^2}{2\alpha^2}\right)}$$

Where: (1) α represents the standard deviation of the Gaussian kernel; (2) x_i represents the rank of the ith ranked blind data instance computed from the AD2 values across all training and blind data points; and (3) x_j represents the rank of the jth ranked training data instance value of the AD2 values across all training and blind data points.

The value determined for the parameter, α, is 61.615. That value was determined by n-fold cross-validation on the training data. The α parameter selected was one that produced the minimal value for -2*log likelihood over the training data, which is the maximum likelihood estimator for these data, assuming Bernoulli errors.

Figure 45 shows the derived probability model plotted against the rankings of the UXO and Not-UXO on the training data. The red circles show the ranks of UXO in the training data. The green show the ranks of Not-UXO in the training data. Note that the rankings are derived from Attribute AD2 and represent the rankings across all training and blind data not assigned to cannot-analyze one or two.

Figure 45. Kernel regression fit between UXO and attribute AD2 on training data

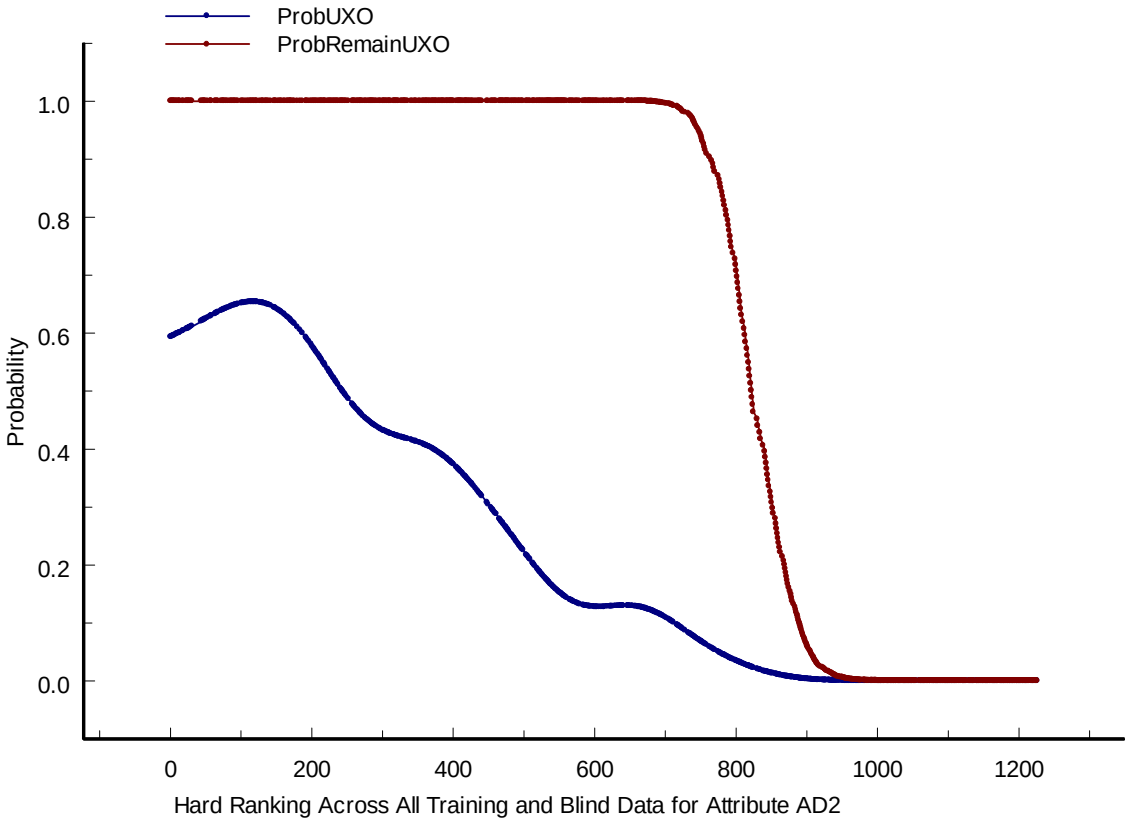


The Gaussian kernel generated by the training data, using the above kernel width parameter, was then applied to the ranked blind data, generating a probability that each blind data item is UXO.

Once individual target probabilities are set, the probability that all blind targets above each AD2 ranking contain one or more UXO is calculated using the approach outlined in 2.1.6. This is the residual risk as a function of rank. In particular, we used Equation 1 and Equation 2 to compute the OR of the probabilities for all targets from the ranking for which the computation is being performed to the most extremely ranked blind target.

Figure 46 shows the result of applying the kernel regression model derived above to the blind data. The blue line in Figure 46 is the probability of UXO as a function of the rank derived from the values of the AD2. The red line is the probability that one or more UXO remain on site to the right of each rank value. When the red line reaches a critical probability value ($p\text{-value}_{crit}$), we assess all targets remaining to the right of that rank (i.e. targets with a larger rank) as high-probability Not-UXO.

Figure 46. Kernel regression applied to blind data



A 95% confidence level was chosen in the experimental plan to set the stop-digging threshold and is used throughout this project. Since there were two risk analyses used to determine two stop-digging thresholds the Bonferonni correction needed to be applied.[29] Using the Bonferonni correction the p-value$_{crit}$ was set to .025 (2.5%).

The p-value$_{crit}$ was then used to determine the critical rank value of 919, inferring that any target with a rank value greater than 919 was high-probability Not-UXO. At that point, the probability of remaining UXO was 0.02499 — in other words, it satisfies the p-value$_{crit}$ criterion.

Table 30 shows the details for the 95% stop-digging threshold.

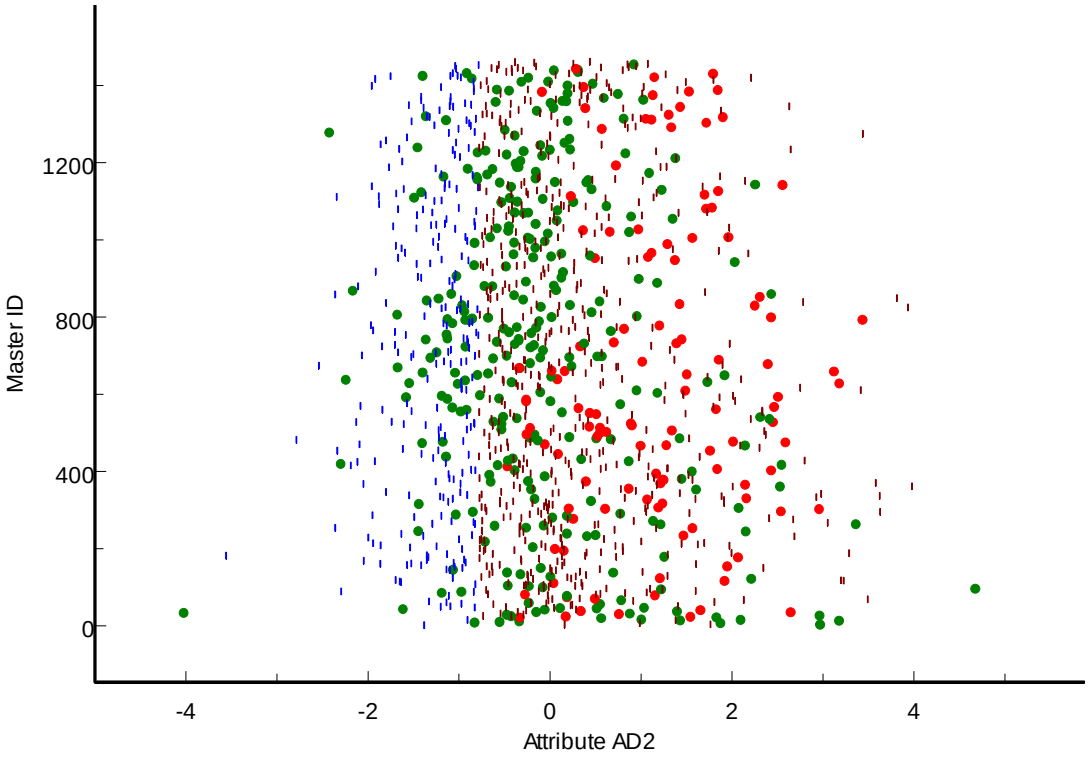
Figure 47 provides a visual representation for which blind targets will be considered safe to leave in the ground.

Table 30. Stop-digging threshold

Iteration 2 AD2						
Confidence	**TID**	**Score**	**HardRank**	**ProbUXO**	**ProbRemainUXO**	**% Of Blind Data Left in Ground**
95.00%	1453	-0.786043025	919	0.00158	0.02499	19.27%

[29] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

Figure 47. Attribute space for attribute AD2 (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data above the stop-digging threshold. Blue Lines are blind data below the stop-digging threshold). Note the y-axis is Master ID which is used to spread out the targets in the graph.



6.11.4 Remove Cannot-Analyze Category Three Targets

Seven additional blind targets were assigned to a cannot-analyze three category because they were blind targets not well defined because of low data density in the training data. Table 31 shows the updated sample sizes after taking into account the targets excluded by the amplitude discriminator and the cannot-analyze three category targets.

Table 31. Targets lost due to amplitude discriminator and cannot-analyze 3 (Low Sample Size)

Iteration 2	Train	Blind	Total
Original Data Sample Size	438	1026	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	413	816	1229
Cannot-Analyze 2 (Outlier)	(1)	0	(1)
Sample Size After Cannot-Analyze 2 Removed	412	816	1228
Amplitude Discriminator Low Prob UXO	(63)	(247)	(310)
Sample Size After Amplitude Discriminator Removed	349	569	918
Cannot-Analyze 3 (Low Sample Size)	0	(7)	(7)
Sample Size After Cannot-Analyze 3 Removed	349	562	911

6.11.5 Attribute Reduction for LGP Modeling

6.11.5.1 Attribute Reduction Tools

For attribute reduction, we used the same set of techniques described in Section 6.9.3.1.

6.11.5.2 Attribute Reduction Process

As in iteration one, the first step was to bin all of the attributes using equal-frequency and Chi-square binning procedures. We then used MRMR to reduce the data set to a twenty-four attribute set.

These attributes were further reduced using a J48 single decision tree algorithm which is an extension of the classic C4.5 decision tree algorithm.[30] The resulting tree was as follows: The integers in this tree represent attributes.

J48 pruned tree limb 1:

3 <= 27.991484

| 19 <= 1.082857

| | 17 <= 1.331744: 0 (41.0)

| | 17 > 1.331744

| | | 4 <= 0.194248: 0 (43.0/4.0)

| | | 4 > 0.194248

| | | | 9 <= 0.127186

| | | | | 14 <= 0.030541: 1 (11.0)

| | | | | 14 > 0.030541

| | | | | | 8 <= 0.412508: 1 (4.0/1.0)

| | | | | | 8 > 0.412508: 0 (3.0)

| | | | 9 > 0.127186: 0 (6.0)

| 19 > 1.082857: 0 (77.0)

J48 pruned tree limb 2:

3 > 27.991484

| 13 <= 1.070938

| | 21 <= 1.197813

| | | 19 <= 1.093993

| | | | 16 <= 0.24174: 1 (102.0/19.0)

[30] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

| | | | 16 > 0.24174

| | | | | 14 <= 0.18341: 0 (5.0)

| | | | | 14 > 0.18341: 1 (3.0/1.0)

| | | 19 > 1.093993

| | | | 6 <= 10.541287

| | | | | 6 <= 6.093486: 0 (9.0)

| | | | | 6 > 6.093486

| | | | | | 15 <= 0.035152

| | | | | | | 2 <= 134.139999: 1 (4.0)

| | | | | | | 2 > 134.139999

| | | | | | | | 1 <= 61.81: 0 (4.0)

| | | | | | | | 1 > 61.81: 1 (2.0)

| | | | | | 15 > 0.035152: 0 (5.0)

| | | | 6 > 10.541287: 1 (8.0)

| | 21 > 1.197813

| | | 3 <= 89.778846: 1 (2.0)

| | | 3 > 89.778846

| | | | 16 <= 0.24174: 0 (14.0)

| | | | 16 > 0.24174: 1 (3.0/1.0)

| 13 > 1.070938: 0 (19.0)

The tree identified five attributes (3, 19, 13, 17, and 21) as the likely most important for classifying UXO. In addition, we add attributes 1 and 2 (from the amplitude discriminator) and then calculate the pseudo-principal components for each pair-wise combination of the seven variables. From there more J48 trees were run and variables from the top of the tree were selected.

The top 4 variables from the J48 trees were then used to run multiple Random Forests[31] models. Random Forests is an ensemble decision tree algorithm that is reasonably fast and is does a good job of building preliminary models. The most favorable Random Forests model resulted from using attributes A, B, C and D. Where:

1. Attribute A = Second pseudo-principal component of the:
    - Base 10 log of the second moment in the inner part of the ellipse for the second decay channel; and

[31] Random Forests™ is a trademark of Leo Breiman.

- First moment of the ratios of the first decay channel and second decay channel in the inner part of the ellipse.

2. Attribute B = Second pseudo-principal component of the:
   - First moment of the ratios of the first/second decay channel and second/third decay channel in the outer part of the ellipse; and
   - Base 10 log of the first moment for the fourth decay channel in the ellipse.
3. Attribute C = First orthogonal component of the:
   - First moment of the ratios of the second/third decay channel and third/fourth decay channel in the ellipse; and
   - First moment of the ratios of the first/second decay channel and second/third decay channel in the outer part of the ellipse.
4. Attribute D = First orthogonal component of the:
   - First moment of the ratios of the first decay channel and second decay channel in the inner part of the ellipse; and
   - First moment of the ratios of the first/second decay channel and second/third decay channel in the outer part of the ellipse.

These are the attributes that were used in the LGP discrimination process. The results from the Random Forests models are:

Random Forest:

Test mode: 50-fold cross-validation

Random forest of 1000 trees, each constructed while considering 2 random features.

Out of bag error: 0.2264

Time taken to build model: 3.04 seconds

Summary:

Correctly Classified Instances	272	77.937 %
Incorrectly Classified Instances	77	22.063 %
Kappa statistic		0.5176
Mean absolute error		0.2741
Root mean squared error		0.3883
Relative absolute error		60.45%
Root relative squared error		81.55 %
Total Number of Instances		349

Detailed Accuracy By Class:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.82	0.298	0.839	0.82	0.829	0.855	0
0.702	0.18	0.675	0.702	0.688	0.855	1
0.779	0.257	0.782	0.779	0.78	0.855	Weighted Avg.

Confusion Matrix:

		Predicted	
		Not-UXO	**UXO**
Actual	**Not-UXO**	187	41
	UXO	36	85

6.11.5.3 Reduced Attributes--Match between Training and Blind Data

Since the training data will be used to make predictions to the blind data we looked at the descriptive statistics and graphs for all four attributes individually to identify how well the training and blind data "match". The descriptive statistics are shown in Table 32, Table 33, Table 34, and Table 35 and the graphics in Figure 48, Figure 49, Figure 50, and Figure 51.

As in the descriptive statistics for the amplitude discriminator for this iteration, these tables and figures show the expected bias that results from the over-sampling of UXO in our request for more groundtruth. That is, we moved a disproportionate number of UXO from blind data to training data. Therefore, attributes that are predictive of UXO will be biased as between the two data sets. Given the foregoing, we believed there was a reasonable degree of consistency between the training and blind data.

Table 32. Descriptive statistics for Attribute A

Attribute A											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	911	-0.09315	0.43356	0.18797	0.56022	-0.16891	3.10739	0.29	0.000*	0.246	>.10
Training Data	349	0.00796	0.41831	0.17498	0.56997	-0.03074	2.50779	0.139	0.284	-0.001	>.10
Blind Data	562	-0.15594	0.43134	0.18605	0.51838	-0.23451	3.10739	0.421	0.000*	0.596	.10-.05

Figure 48. Box and whisker plots for Attribute A

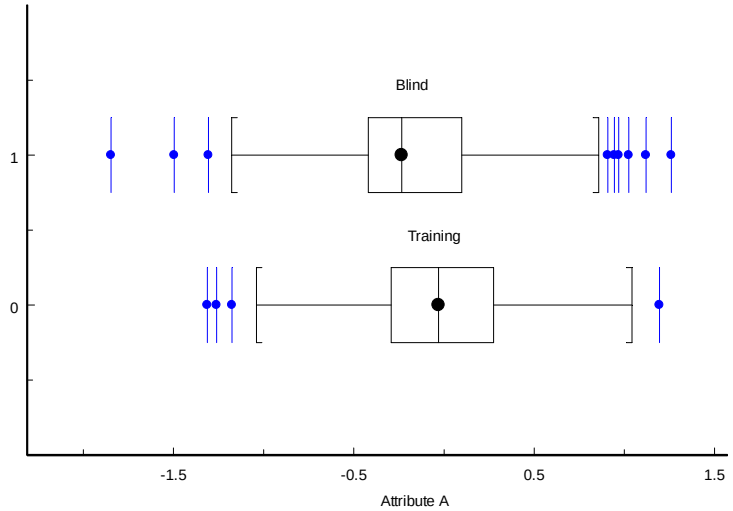


Table 33. Descriptive statistics for Attribute B

Attribute B											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	911	-0.66779	0.05949	0.00354	0.07157	-0.67344	0.42391	0.935	0.000*	2.308	<.02*
Training Data	349	-0.67933	0.05156	0.00266	0.06193	-0.68305	0.37926	0.847	0.000*	2.529	<.02*
Blind Data	562	-0.66062	0.0629	0.00396	0.07557	-0.66541	0.42353	0.886	0.000*	2.009	<.02*

Figure 49. Box and whisker plots for Attribute B

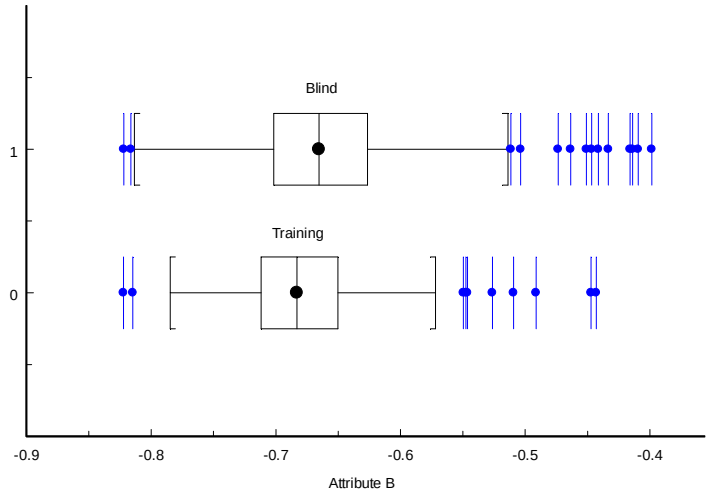


Table 34. Descriptive statistics for Attribute C

Attribute C											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	911	2.14683	0.06242	0.0039	0.06658	2.13493	0.48734	1.623	0.000*	4.436	<.02*
Training Data	349	2.13616	0.05068	0.00257	0.05609	2.12777	0.3636	1.365	0.000*	3.625	<.02*
Blind Data	562	2.15346	0.06791	0.00461	0.08059	2.13801	0.48734	1.567	0.000*	3.864	<.02*

Figure 50. Box and whisker plots for Attribute C

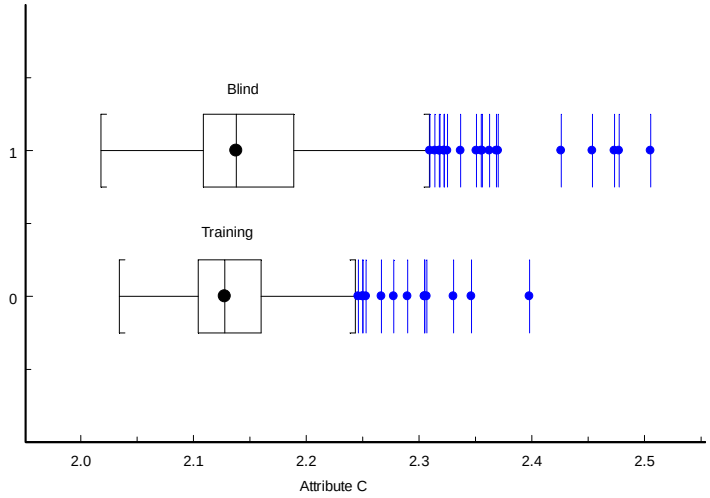
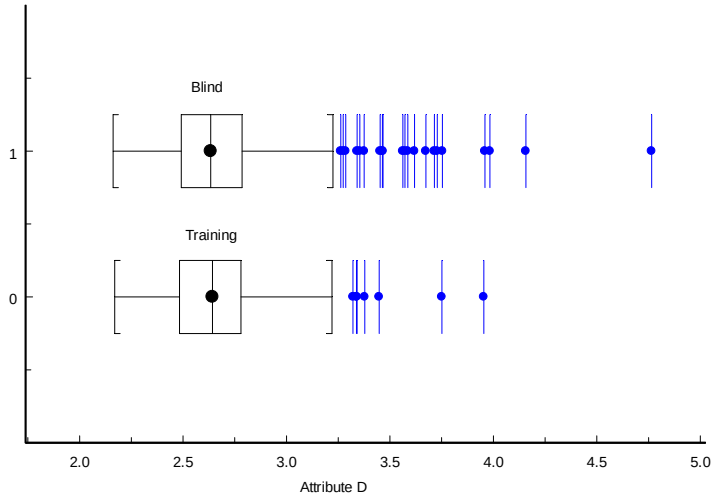


Table 35. Descriptive statistics for Attribute D

Attribute D											
Data	**Sample Size**	**Mean**	**Std Dev**	**Variance**	**IQR**	**Median**	**Range**	**Skewness**	**p-value**	**Kurtosis**	**p-value**
All Data	911	2.66345	0.26259	0.06895	0.29487	2.63571	2.60372	1.938	0.000*	8.535	<.02*
Training Data	349	2.6489	0.23012	0.05295	0.30029	2.64229	1.78389	1.154	0.000*	4.236	<.02*
Blind Data	562	2.67248	0.28069	0.07879	0.29405	2.63435	2.60372	2.152	0.000*	9.155	<.02*

Figure 51. Box and whisker plots for Attribute D

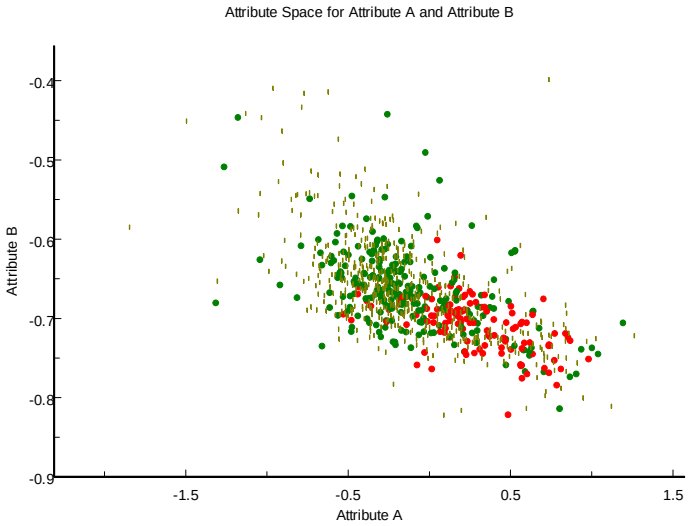


6.11.5.4 Attribute Space Graphs

The graphs in this section show the attribute space for the four selected attributes. In these graphs, red circles represent UXO in the training data, green circles represent Not-UXO in the training data and brown lines represent the blind data.

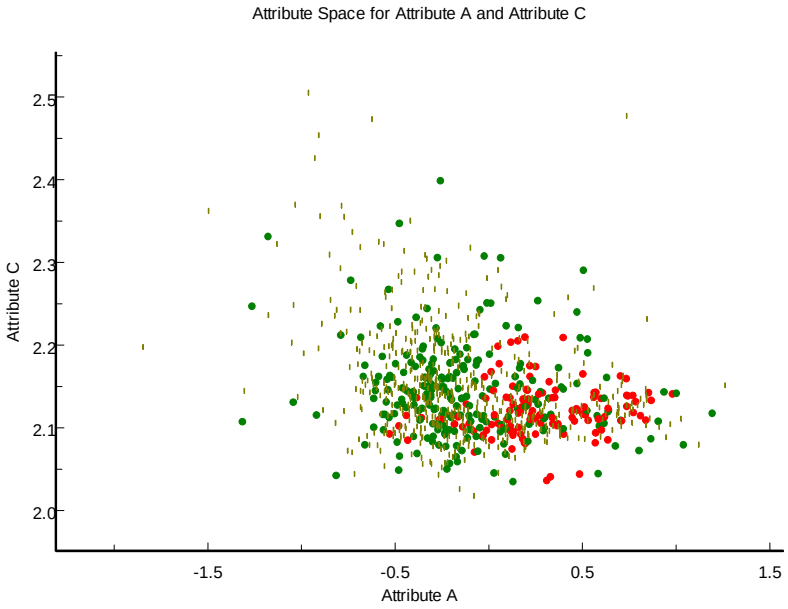
6.11.5.4.1 Attribute A versus Attribute B

Figure 52. Attribute space for attribute A versus attribute B (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data).



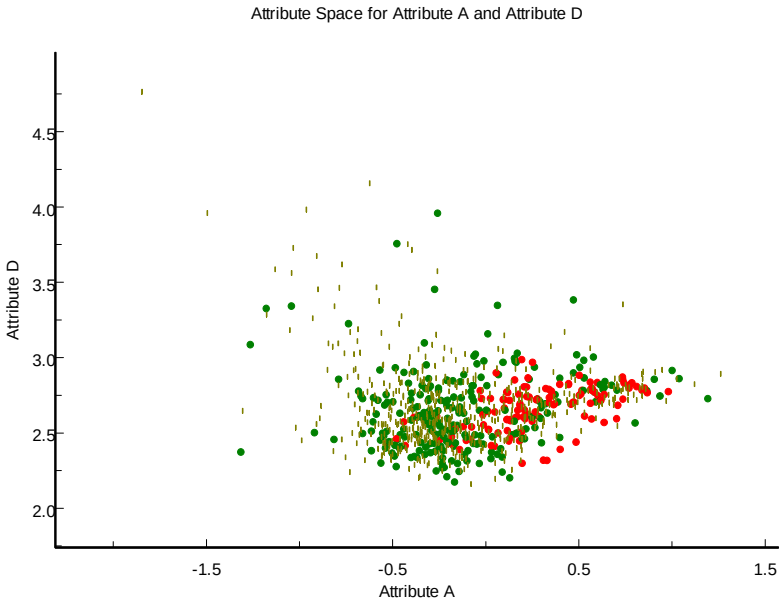
6.11.5.4.2 Attribute A versus Attribute C

Figure 53. Attribute space for attribute A versus attribute C. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data).



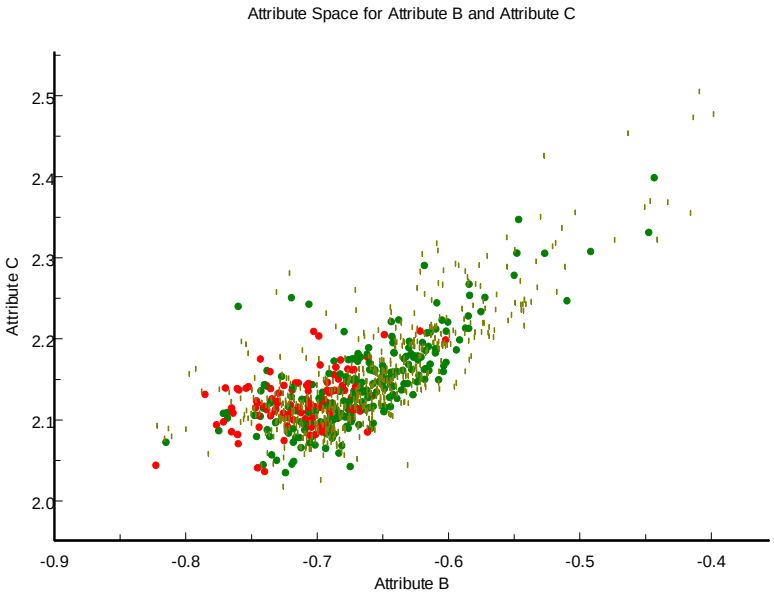
6.11.5.4.3 Attribute A versus Attribute D

Figure 54. Attribute space for attribute A versus attribute D. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data).



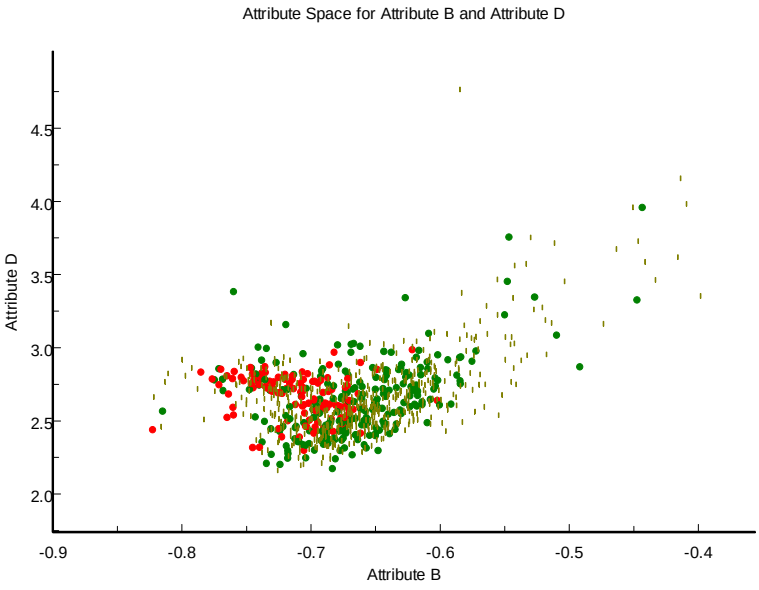
6.11.5.4.4 Attribute B versus Attribute C

Figure 55. Attribute space for attribute B versus attribute C. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data).



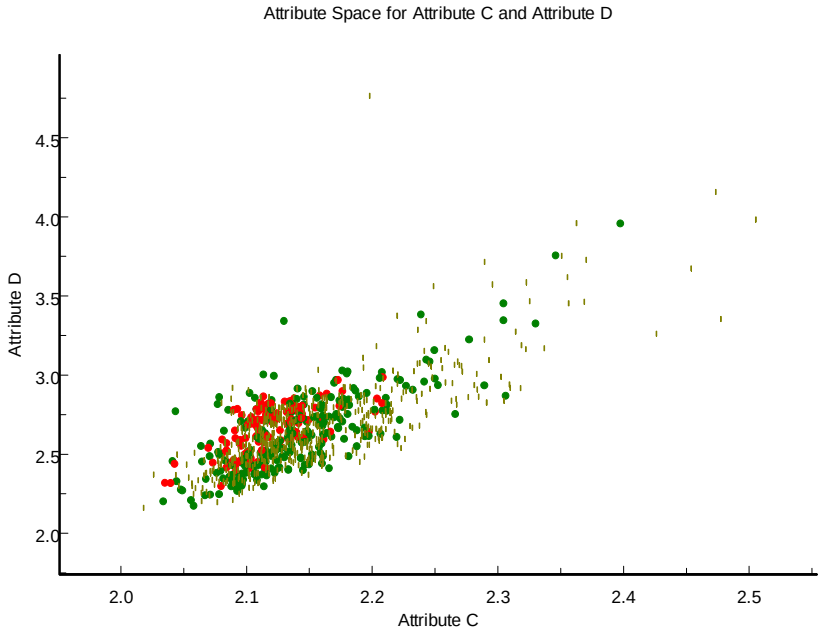
6.11.5.4.5 Attribute B versus Attribute D

Figure 56. Attribute space for attribute B versus attribute D. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data).



6.11.5.4.6 Attribute C versus Attribute D

Figure 57. Attribute space for Attribute C versus attribute D. (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data).



6.11.5.5 Remove Cannot-Analyze Category Four Targets

At this point, we reviewed the attribute space graphs and removed blind targets that were extreme outliers in multi-dimensional attribute space. In looking at the attribute space graphs we decided to send two blind targets to cannot-analyze four because they appeared to be outliers in attribute space. Figure 57 shows an example of one of those targets. It is the small brown point at the top of the chart, coordinates are approximately (2.2, 5.0).

Table 36 shows the updated sample sizes after taking into account cannot-analyze four.

Table 36. Count of targets lost due to cannot-analyze 4 (Outliers)

Iteration 2	Train	Blind	Total
Original Data Sample Size	438	1026	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	413	816	1229
Cannot-Analyze 2 (Outlier)	(1)	0	(1)
Sample Size After Cannot-Analyze 2 Removed	412	816	1228
Amplitude Discriminator Low Prob UXO	(63)	(247)	(310)
Sample Size After Amplitude Discriminator Removed	349	569	918
Cannot-Analyze 3 (Low Sample Size)	0	(7)	(7)
Sample Size After Cannot-Analyze 3 Removed	349	562	911
Cannot-Analyze 4 (Outliers)	0	(2)	(2)
Sample Size After Cannot-Analyze 4 Removed	349	560	909

6.11.6 LGP Discriminator

6.11.6.1 Training Data Used

Discrimination was performed using Discipulus LGP. We used all training targets not assigned to cannot-analyze categories one, two, three or four and not assessed as high-probability Not-UXO by the amplitude discriminator. That is summarized in Table 36.

6.11.6.2 Parameters used for Deriving LGP Model on the Training Data

6.11.6.2.1 LGP Software Used

Discipulus™ 5.0 was used to generate our model on the remaining training data using the four attributes (A, B, C, and D) as our input variables (independent variables) and whether or not a UXO was found as our output variable (dependent variable).

6.11.6.2.2 Best Program Selection Procedure

At the end of each Discipulus project, we opened the program designated by Discipulus as the best program of the project and we repeatedly removed introns from that program until the best program ceased getting shorter. The best program with introns removed was selected as the program model for that Discipulus project.

6.11.6.2.3 Fitness Function Used

The fitness function used for all LGP modeling runs was "Ranking-Best ROC Curve" as measured by area underneath the curve.

6.11.6.2.4 Parameter Settings for Discipulus™ 5.0 Runs

The parameter settings for all LGP modeling runs are the default Discipulus™ 5.0 parameters with the following changes:

- Stepping = Disabled
- Single Run Termination:
  - Generations without Improvement = 225
  - Use Adaptive Termination = Disabled
- Batch Run Termination:
  - Maximum Number of Runs = 20
- Single Run Parameters:
  - Population Size = No Randomization; Set to 2000
  - Maximum Program Size = No Randomization; Set to 256
  - Subset Size = No Randomization; Set to 100

6.11.6.2.5 Cross-Validation Runs to Determine Noise Parameter

We added noise to our independent variables in our LGP projects as described in Section 6.9.4.2. In order to determine how much noise to add to our independent variables we ran LGP using

cross-validation for varying noise levels from 6% to 12% (6%, 7%, 8%....12%). The LGP input data for each noise level was created using the following settings:

- Number of Cross Validation Folds = 20
- Multiplication Factor = 30

This resulted in 7 different models—one for each noise level test. Out of the 7 models the best models were chosen using two criteria:

1. Best ROC curve as measured by area underneath the curve
2. Count of Not-UXO data remaining after last UXO was "found"

Table 37. Cross-Validition errors for different noise levels

Iteration 2		
Noise	AUC	Misranked
7%	86.38%	175
8%	84.25%	134
9%	84.91%	176
10%	85.46%	165
11%	85.73%	210

Using these two metrics we determined that the noise level at 8% would likely produce the best model. Note the trade-off that exists here between higher area under the curve (7%) and superior location of the last UXO (8%, measured by misranked). We deliberately selected the better final location setting.

6.11.6.3 Creating the LGP Ensemble Predictor

Using the 8% noise level we conducted thirty LGP bagging projects with the following settings: The setup and logic of the bagging process is described in Section 6.9.4.3.

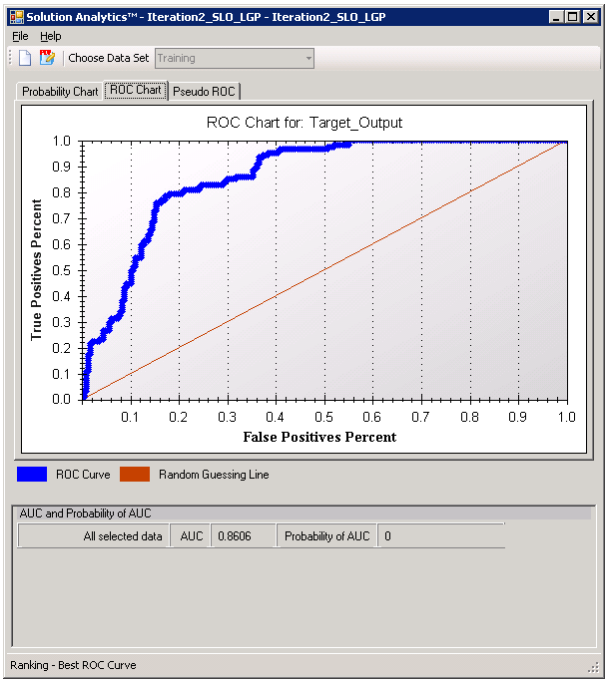
Each bag was modeled by LGP and the average LGP output of these 30 models was used to make ensemble predictions for each target.

The predictions from the 8% model will be used to generate predictions on the blind data and will be referred to as the "LGP ensemble predictor."

6.11.6.4 Predictive Error on the Training Data

The area under the curve of the ROC curve on the training data of the derived model was 86.06%. These results are shown only on the out-of-bag training data summed across all bags. Out-of-bag data is not used to train the model. Figure 58 shows the ROC curve generated by the derived LGP model on the data used to train the LGP ensemble predictor.

Figure 58. ROC curve on training data for LGP model



6.11.6.5 Assign Attribute Space Outliers to Cannot-Analyze Five Category

Before proceeding further, we again reviewed attribute space and determined to assign twelve additional blind targets the cannot-analyze category five. They appeared to be outliers in multi dimensional attribute space. Figure 59 shows an example of the process. The reddish circle delineates five blind targets we adjudged too far from any training examples to make a decision.

Figure 59. Examples of targets assigned to cannot-analyze five category as attribute space outliers.

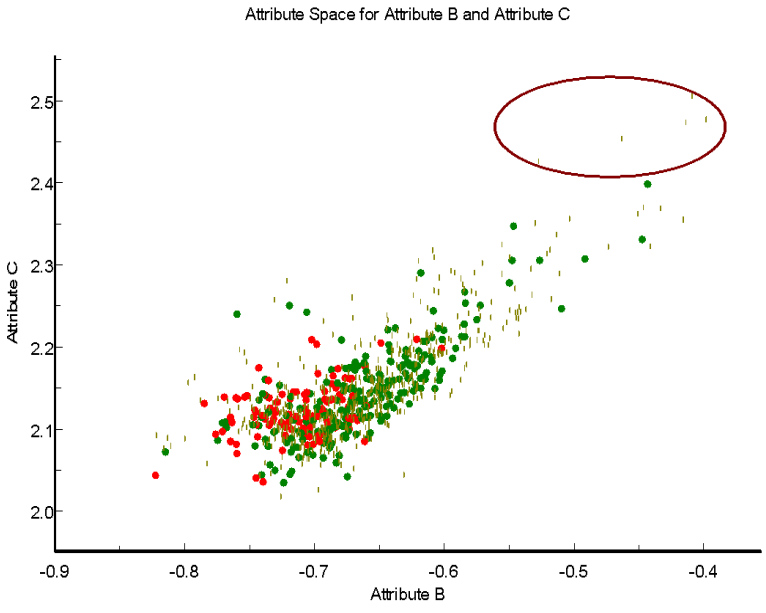


Table 38 shows the updated sample sizes after taking into account cannot-analyze five category targets.

Table 38. Count of targets lost due to cannot-analyze 5 (Outliers)

Iteration 2	Train	Blind	Total
Original Data Sample Size	438	1026	1464
Cannot-Analyze 1 (Visual Inspection)	(25)	(210)	(235)
Sample Size After Cannot-Analyze 1 Removed	413	816	1229
Cannot-Analyze 2 (Outlier)	(1)	0	(1)
Sample Size After Cannot-Analyze 2 Removed	412	816	1228
Amplitude Discriminator Low Prob UXO	(63)	(247)	(310)
Sample Size After Amplitude Discriminator Removed	349	569	918
Cannot-Analyze 3 (Low Sample Size)	0	(7)	(7)
Sample Size After Cannot-Analyze 3 Removed	349	562	911
Cannot-Analyze 4 (Outliers)	0	(2)	(2)
Sample Size After Cannot-Analyze 4 Removed	349	560	909
Cannot-Analyze 5 (Outliers)	(1)	(12)	(13)
Sample Size After Cannot-Analyze 5 Removed	348	548	896

6.11.7 Risk Analysis/Stop-Digging Threshold

The next step in this process is to determine the LGP stop-digging threshold. The LGP values were first converted into ranks across the entire training and blind data. Lower LGP values were interpreted as larger rankings. Next, kernel regression with a Gaussian kernel was used to determine the Probability of UXO for each target:

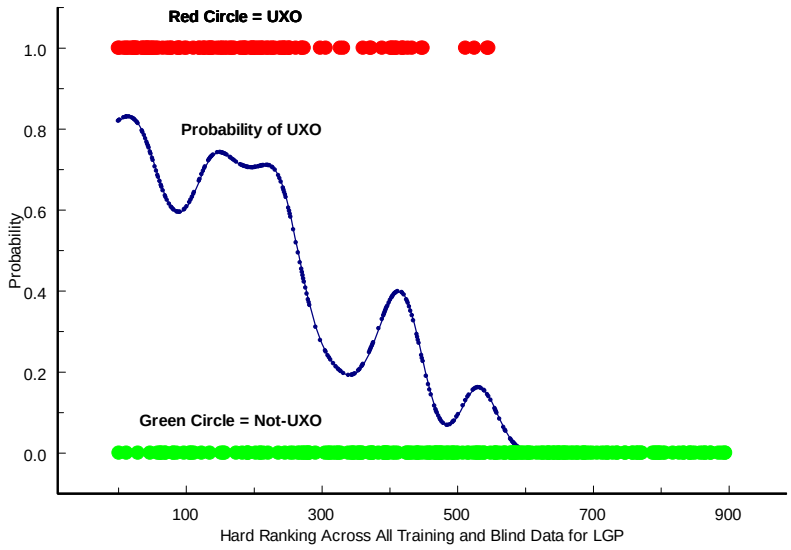
$$\mathrm{P}(UXO)_i = \sum_j e^{-\left(\frac{(x_i - x_j)^2}{2\alpha^2}\right)}$$

Where: (1) α represents the standard deviation of the Gaussian kernel; (2) x_i represents rank of the *ith* ranked blind data instance computed from the LGP values across all training and blind data points; and (3) x_j represents rank of the *jth* ranked training data instance value of the LGP values across all training and blind data points.

The value determined for the parameter, α, is 23.001. That value was determined by n-fold cross-validation on the training data. The α parameter selected was one that produced the minimal value for -2*log likelihood over the training data, which is the maximum likelihood estimator for these data, assuming Bernoulli errors.

Figure 60 shows the derived model plotted against the rankings of the UXO and Not-UXO on the training data. Note that the rankings are derived from the LGP values and represent the rankings across all training and blind data not assigned to cannot-analyze one, 2, 3, 4, 5 or that were removed by the amplitude discriminator.

Figure 60. Kernel regression fit between UXO and LGP on training data

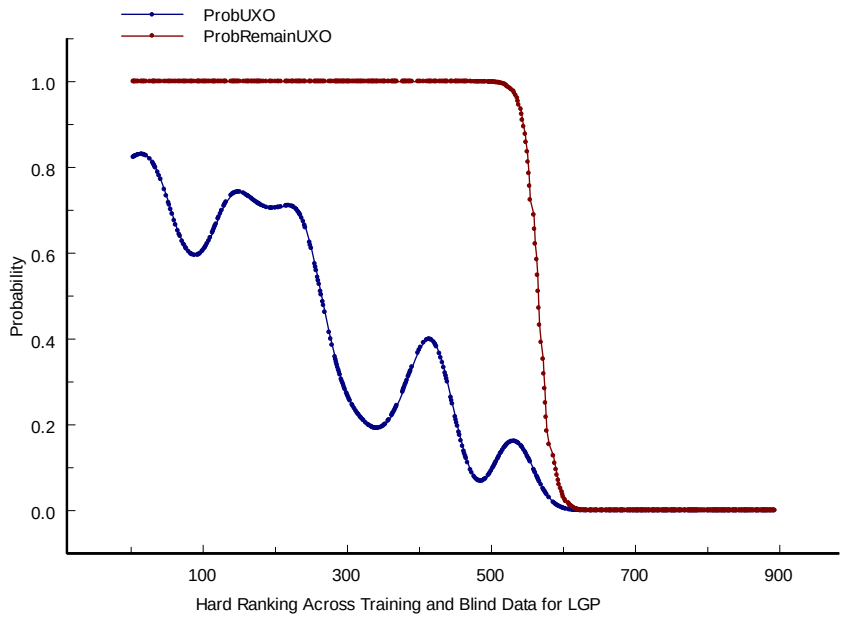


The Gaussian kernel width generated by the training data was then applied to the ranked blind data, generating a probability that each blind data item is UXO.

Once individual target probabilities are set, the probability that all blind targets above each ranking contain one or more UXO is calculated using the approach outlined in 2.1.6. This is the residual risk as a function of rank. In particular, we used Equation 1 and Equation 2 to compute the OR of the probabilities for all targets from the ranking for which the computation is being performed to the most extremely ranked blind target.

Figure 61 shows the result of applying the kernel model defined above to the blind data. The blue line in Figure 61 is the probability of UXO as a function of the rank derived from the LGP values. The red line is the probability that one or more UXO remain on site at each rank value to the right of that rank. When the red line reaches a critical probability value ($p\text{-value}_{crit}$), we assess all targets remaining to the right of that rank (i.e. targets with a larger rank) as high-probability Not-UXO.

Figure 61. Kernel regression applied to blind data



A 95% confidence level was designated in the experimental plan and is used throughout this project. Since there were two risk analyses used to determine two stop-digging thresholds the Bonferonni correction needed to be applied.[32] Using the Bonferonni correction the $p\text{-value}_{crit}$ was set to .025 (2.5%).

The $p\text{-value}_{crit}$ was then used to determine the critical rank stop-digging value of 603, inferring that any target with a rank value greater than 603 was high-probability Not-UXO. At that point, the probability of remaining UXO was 0.02133—in other words, it satisfies the $p\text{-value}_{crit}$ criterion.

Table 39 shows the details for the 95% stop-digging threshold.

Table 39. Stop-digging threshold

Iteration 2 LGP					
Confidence	**TID**	**HardRank**	**ProbUXO**	**ProbRemainUXO**	**% Of Blind Data Left in Ground**
95.00%	319	603	0.00401	0.02133	16.61%

6.11.8 Prepare Prioritized Dig-List

The blind data targets that fell below the amplitude discriminator stop-digging threshold were combined with the blind data targets that fell below LGP discriminator stop-digging threshold. This 'final below threshold list' contains the blind data targets identified as safe to leave in the ground. We then calculated the total % of blind data left in the ground by summing the percentage left in the ground using the amplitude discriminator (19.27%) and the percentage left in the ground using the LGP discriminator (16.61%). This results in 35.88% targets out of the original 1282 blind targets that may be left in the ground.

[32] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

To create the prioritized dig list we added the blind data targets that fell above LGP discriminator stop-digging threshold to the 'final below threshold list' and re-ranked the targets based on Probability of UXO (note this was a slightly different procedure from iteration one). Finally, the targets assessed as cannot-analyze were added to the list creating our final prioritized dig list.

6.11.9 Attribute Importance Analysis

Table 40 summarizes the importance of the four attributes described above in solving the UXO/Not-UXO ranking problem for iteration two. These values are averaged over all bagging projects that formed the final LGP ensemble predictor. The "frequency" column measures what proportion of best programs in all thirty projects contained that attribute as an input in the program. The "maximum" column shows the maximum impact that variable had on the area under the curve over all best programs of all thirty bags. Thus, Attribute A increased the area under the curve produced by one of the thirty programs by 0.137. The "average" column shows the average amount by which the attribute increased the fitness over all thirty best programs. That means that on average, Attribute A increased the AUC of all best programs by 0.085.

Table 40. LGP attribute importance analysis

	IMPACTS 8%		
Attribute	**Frequency**	**Maximum**	**Average**
Attribute A	1	0.136736	0.085096
Attribute B	1	0.060083	0.031745
Attribute C	0.9445	0.031092	0.012013
Attribute D	1	0.068843	0.052372

There are three conclusions from these data:

1. All four of the selected attributes were highly relevant to solving the ranking problem;
2. Each of them contributed significantly to increasing area under the curve. Attribute A was by far the strongest contributor; and
3. Because all attributes occur with high frequency, all four attributes likely contain significant amounts of information about solving the ranking problem that is different than the others. If these attributes contained just duplicate information, one would expect to see different best programs substitute them for each other and thus, a lower frequency for the attributes with duplicate information.

7 PERFORMANCE ASSESSMENT

7.1 INTRODUCTION

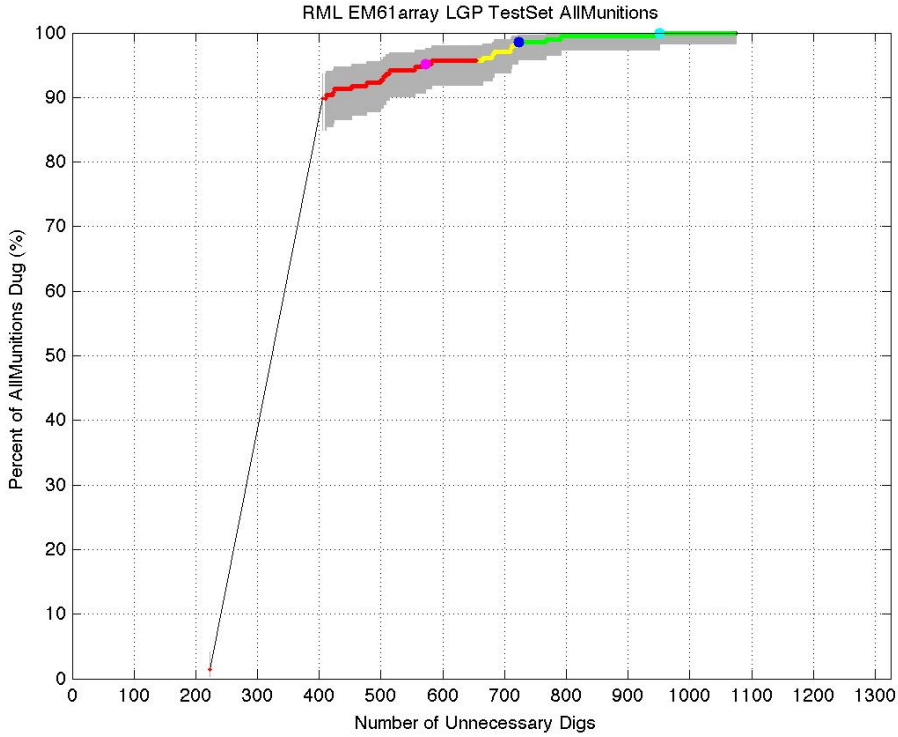
The presentation in this section will be to briefly present the result against the performance objectives and then treat, in greater detail, the areas in which we did not meet those objectives.

The following figures and data will be used in the discussion of more than one of our objectives. Accordingly, they are presented in this introductory section in sections corresponding to the two iterations. In terms of their importance, iteration two is the more significant as it more closely resembles the rankings that would be generated by the LGP Discrimination Process on a real site cleanup and comprised our final dig-list.

7.1.1 Iteration One Overview

After we had submitted our prioritized dig-list for both iterations, the Program Office returned ground truth and a graphic of the performance of each dig-list as a ROC chart. Figure 62 is that ROC chart for iteration one.

Figure 62. ROC curve on blind data for iteration one prioritized dig list



In this figure, the gray line starts at approximately 220 on the x-axis. That represents all cannot-analyze targets for this iteration. The gray line represents the top-ranked targets on our dig list. They were tied for "first-place." What the gray line indicates is that in the first 180 targets on our dig-list, we located 90% of the UXO. The dark blue circle is the point at which we set the stop digging threshold and the green line is all targets below the stop-digging threshold. The final UXO was located at the light blue circle at about ranking 950 on the x-axis.

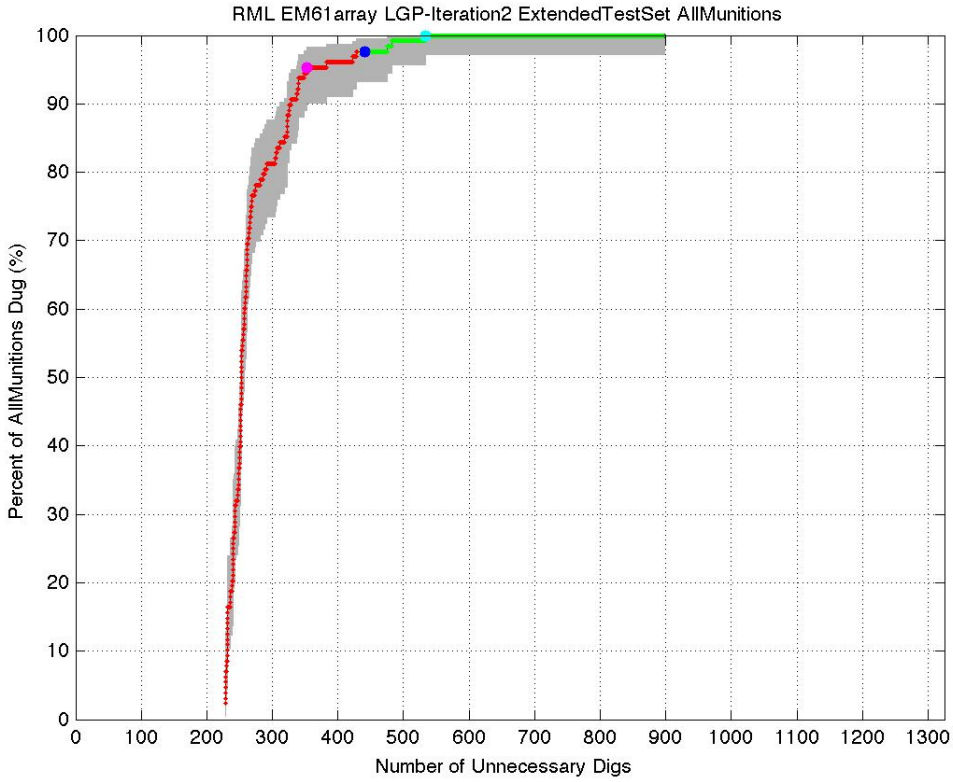
The areas under the curve for this ROC chart may be measured in two ways.

1. Including the cannot-analyze targets, the area under the curve is 0.683
2. Including only targets we ranked with our discriminators, the area under the curve is 0.858.

7.1.2 Iteration Two Overview

After we had submitted our prioritized dig-list for both iterations, the Program Office returned ground truth and a graphic of the performance of the iteration two dig-list as a ROC chart. Figure 63 is that ROC chart.

Figure 63. ROC curve on blind data for iteration two dig-list



In this figure, each red dot represents a UXO located on our dig-list. The first one is shown at approximately 220 on the x-axis. That gap before 220 represents all cannot-analyze targets for this iteration. This chart shows that we located 90% of the UXO in the first 100 targets ranked by our LGP ensemble predictor or the amplitude discriminator. The dark blue circle in this figure is the point at which we set the stop-digging threshold and the green line represents all targets below the stop-digging threshold. The final UXO was located at the light blue circle at about ranking 540 on the x-axis.

The areas under the curve for this ROC chart may be measured in two ways.

1. Including the cannot-analyze targets, the area under the curve is 0.703.
2. Including only targets we ranked with our discriminators, the area under the curve is 0.936.

7.2 OBJECTIVE: MAXIMIZE CORRECT CLASSIFICATION OF MUNITIONS

7.2.1 Introduction

On both iterations one and two, 98.6% of UXO were ranked above the stop-digging threshold. Our objective was to rank 100% above the stop-digging threshold. Put another way, three UXO were ranked below the stop-digging threshold (false negatives) on both iterations. Two of the targets (target 16 and 1444) were below the stop digging threshold on both iterations. Altogether,

four targets appeared below the stop digging threshold on at least one iteration dig-list. On both iterations, target 1444 was the final UXO identified by our dig-lists.

Table 41 and Table 42 summarize the Master ID's and physical details about these four targets (note that there might be an error in the length for target 1444). Three of the targets missed were 60mm Mortars and one was a 60mm Boom and a 2.36" Rocket. Depth is in centimeter and dip angle is in degrees from horizontal. N/A means "not available."

Table 41. Iteration one false negative target ground-truth

Master ID	Description	Depth	Dip	Dip Angle	Azim	Length
16	60mm Mortar	42	Down	80	42	13
512	60mm Mortar	34	Below	8	167	13
1444	60mm Mortar	32	Below	28	197	27

All iteration one false negatives were generated by the LGP ensemble predictor. That is, they were not excluded by the amplitude discriminator.

Table 42. Iteration two false negative target ground-truth

Master ID	Description	Depth	Dip	Dip Angle	Azim	Length
16	60mm Mortar	42	Down	80	42	13
444	60mm Boom; 2.36" Rocket	78	N/A	N/A	N/A	5
1444	60mm Mortar	32	Below	28	197	27

All iteration two false negatives were excluded as high-probability Not-UXO by the amplitude discriminator, not by the LGP ensemble predictor.

The following four paragraphs summarize our findings about why the four false negatives appeared in this project:

1. Target 1444 (iteration one and two false negative). Target 1444 would have been a false negative under any realistic risk-analysis alternative we investigated. It is a somewhat deep, small 60mm mortar. But other similar 60's were properly classified at similar depths and inclinations. It is, however an extreme outlier in attribute space, even when compared only with other small 60's at similar depths and dip-angles.

2. Target 444 (iteration two false negative only). Target 444 was a mistake. It is a badly overlapped target in a blob and is described in the ground-truth as an very deep (78 cm) 2.36 inch rocket and a 60mm boom. Our notes and our retrospective visual analysis of the DGM indicate it was to be assigned to cannot-analyze; but it was not marked as such in the database. That said, with a slight modification to our risk analysis procedures, it is still properly classified on both iterations.

3. Target 512 (iteration one false negative only). Target 512 was a small, somewhat deep, 60 mm mortar. With a slight modification to our risk analysis procedures, it would have been properly classified on both iterations. It was correctly classified on our final, iteration two dig list.

4. Target 16 (iteration one and two false negative). Target 16 was a small, deep 60 mm mortar but at a relatively favorable dip angle. It was at outlier in attribute space, even when compared with other small, deep 60mm mortars. With a slight modification to our risk analysis procedures, it would have been properly classified on both iterations.

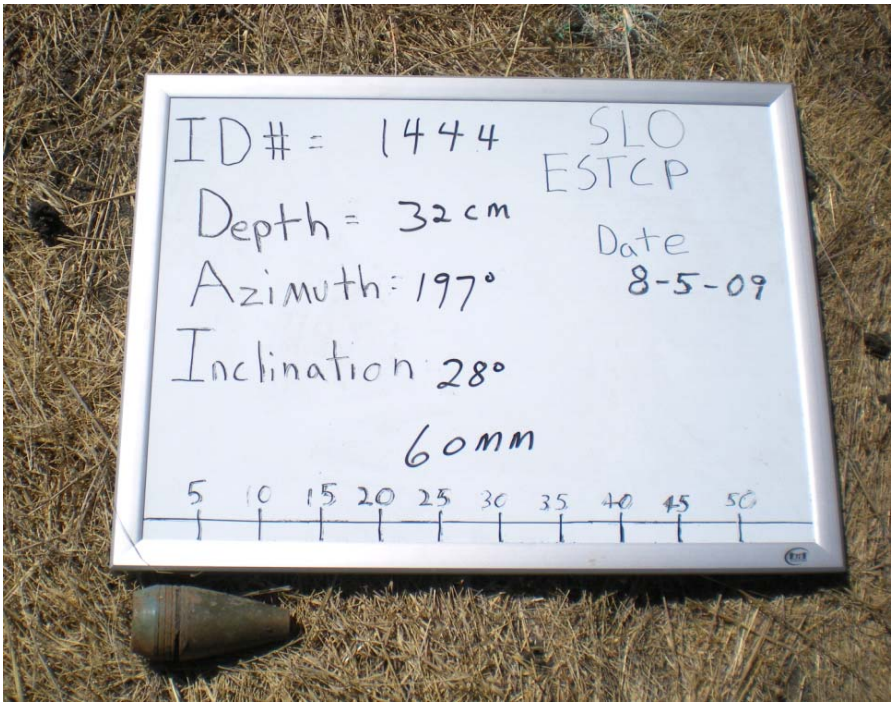
One fairly obvious theme here is that the missed items were either a mistake (444) or small, deep 60mm mortars (512, 16, 1444). By small 60mm mortars, we mean just the head of the mortar with no tail-boom or debris found with it.

7.2.2 Ground-truth, DGM, and defined ellipses of false negatives

7.2.2.1 Target 1444

Target 1444 was a false negative on both iterations. Target 1444 is shown in Figure 64.

Figure 64. Field photo of Target 1444



Two representations of the DGM and the polygon and ellipse we used to define Target 1444 appear below as Figure 65 and Figure 66.

Figure 65 was produced on the data from Oasis Montaj using linear scaling, Channel 1, a minimum millivolt setting of 0 and a maximum millivolt setting of 10. It shows our defined polygon (the geometric basis for our ellipses).

Figure 65. Gridded DGM for Target 1444

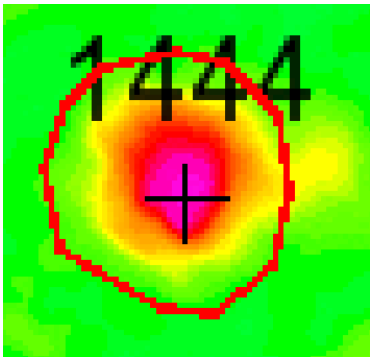
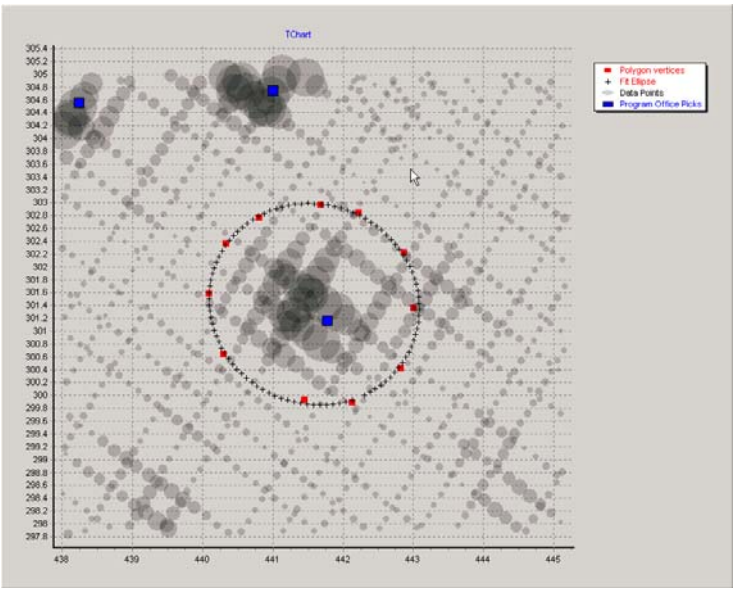


Figure 66 shows the raw data transects for Master ID 1444. Each point represents a single DGM reading for channel one. The size of the point is linearly proportional to the millivolt reading for that point. The minimum millivolt setting was -20 and the maximum millivolt setting was 100. The program office picks are represented by blue squares. The ellipse is our defined target ellipse.

Figure 66. Bubble representation of DGM for Target 1444

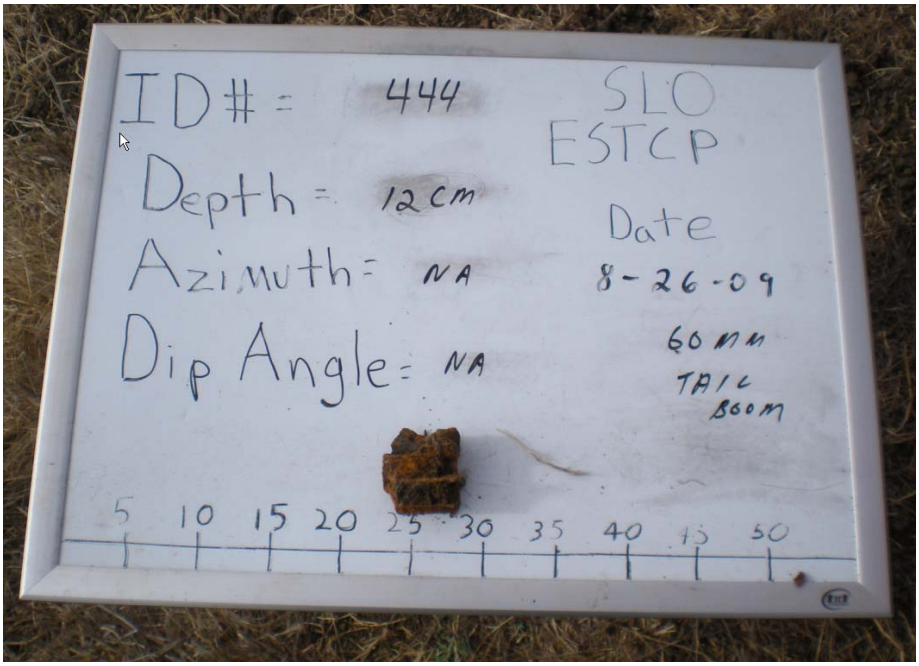


In retrospect, we could quibble about whether the area to the right of the ellipse should have been included. But we would draw the same ellipse again. So there is nothing apparently odd or unusual about the ellipse definition or the field photo that would explain the false negative.

7.2.2.2 Target 444

Target 444 was a false negative only on iteration two. Its field photo is shown in Figure 67.Field photo of Target 444

Figure 67.Field photo of Target 444



Obviously, this item is not UXO. The field photo shows only the tail-boom referenced in the ground-truth lists provided to us by the program office. Some investigation suggested that the 2.36 inch rocket referred to by the program office for this target was originally assigned to Target 435, which overlaps Target 444. This target bears "435" as its field photo id. But the photo was redesignated Target 444a in the ground-truth. Figure 67 shows this redesignated excavated item, which is UXO.

Figure 68.Field photo of Target 435, redesignated as Target 444a

Targets 435 and 444 produced Figure 69 on the gridded data from Oasis Montaj using linear scaling, Channel 1, a minimum millivolt setting of 0 and a maximum millivolt setting of 10.

Figure 69. Master ID 444

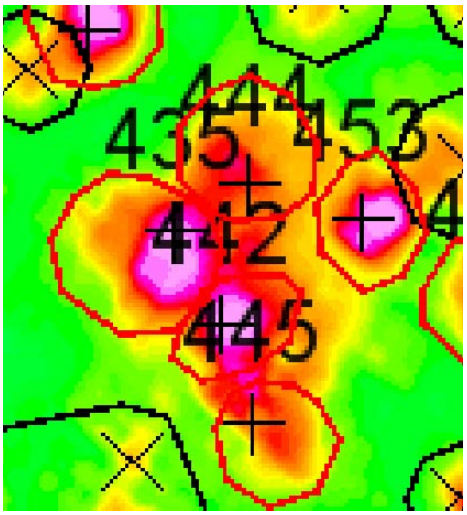
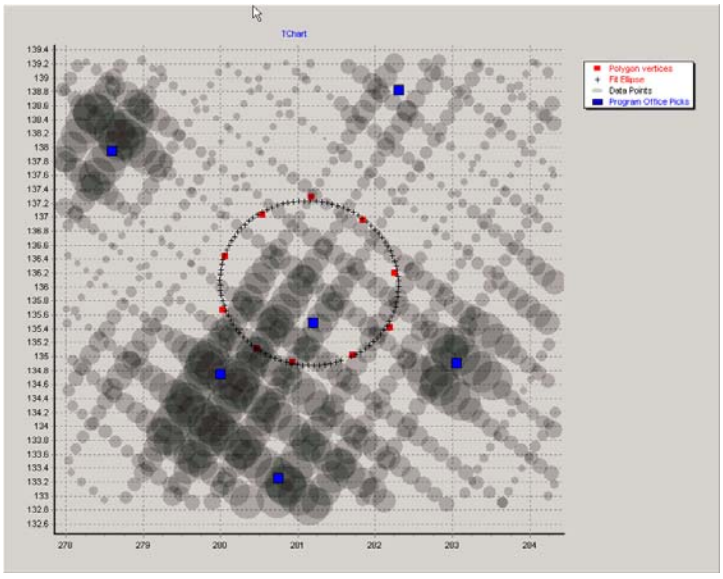


Figure 70 shows the raw data transects for Master ID 444. Each point represents a single DGM reading for a single channel. The size of the point is linearly proportional to the millivolt reading for that point and the minimum millivolt setting was -20 and the maximum millivolt setting was 100.

The Program Office picks are represented by blue squares.

Figure 70. Master ID 444



This is an unusual target for two reasons:

To begin with, it is clear that there was some ambiguity about which Master ID to assign the rocket to so it is not clear precisely what we were attempting to discriminate.

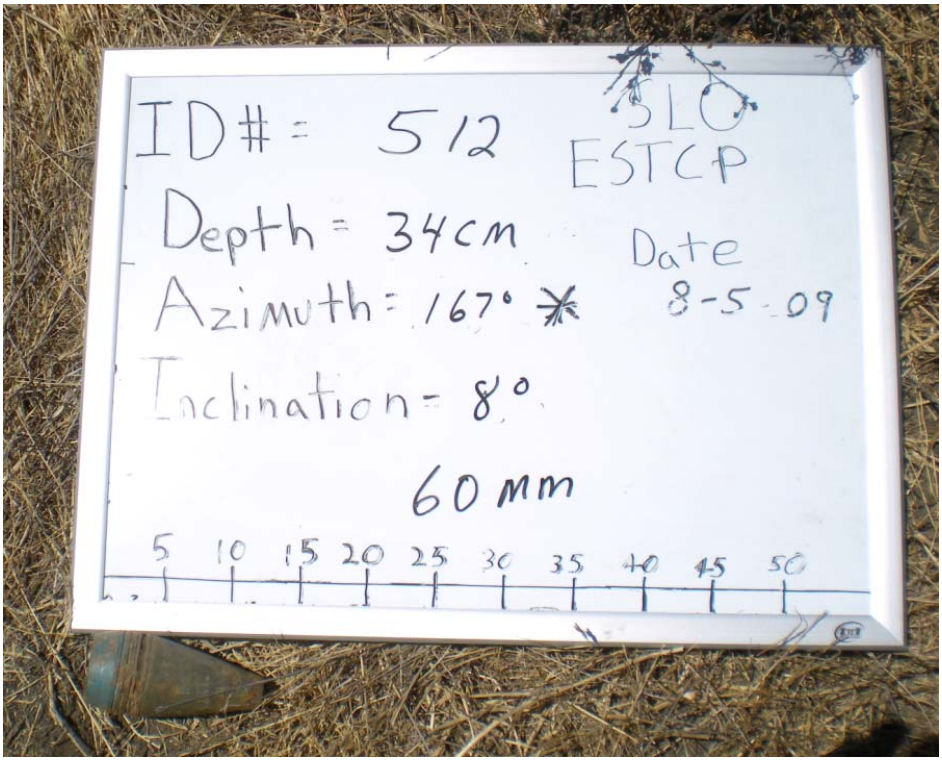
In addition, we made a mistake. As these two figures show, Targets 435, 442, 444, and 445 are part of an overlapping blob and all should have been sent to cannot-analyze category one on both

iterations (See the discussion at Section 6.5.1.1). In fact, our notes indicate that should have occurred and our retrospective analysis does not change that opinion. We do not understand why the cannot-analyze designation for these targets did not get into our database; but it did not. Targets 442 and 445 from the same blob were correctly placed into cannot-analyze category one.

7.2.2.3 Target 512

Target 512 was a false negative on iteration one only. It is shown in Figure 71.

Figure 71. Field photo of Target 512



Two representations of the DGM and the polygon and ellipse we used to define Target 512 appear below as Figure 72 and Figure 73.

Figure 72 was produced on the data from Oasis Montaj using linear scaling, Channel 1, a minimum millivolt setting of 0 and a maximum millivolt setting of 10. It shows our defined polygon (the geometric basis for our ellipses).

Figure 72. Gridded DGM for Target 512

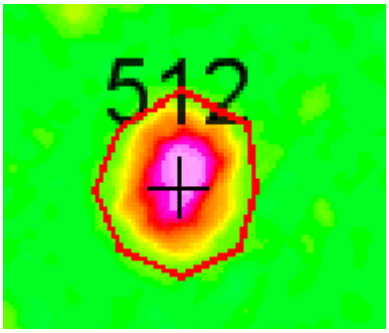
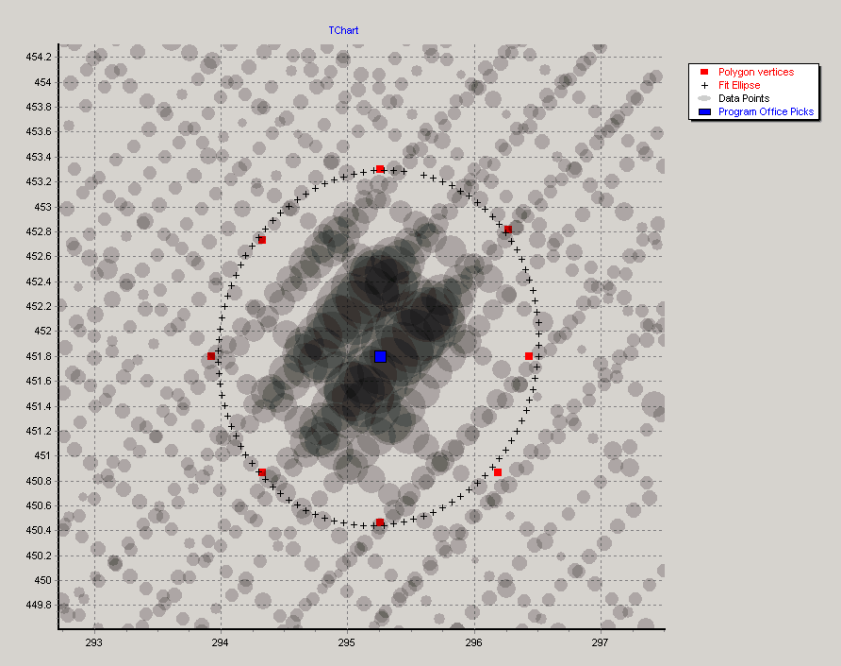


Figure 73 shows the raw data transects for Master ID 512. Each point represents a single DGM reading for channel 1. The size of the point is linearly proportional to the millivolt reading for that point. In this figure, the minimum millivolt setting was -20 and the maximum millivolt setting was 200. The program office picks are represented by blue squares. The ellipse is our defined target ellipse.

Figure 73. Bubble representation of DGM for Target 512



This is a well-defined target with a well-defined ellipse with sufficient data. Our retrospective conclusion is that this ellipse was properly defined and there was sufficient data density to analyze this target. Accordingly, we find no explanation in this part of our process for the misclassification on iteration one.

7.2.2.4 Target 16

Target 16 was a false negative on both iterations. There is no field photo of Target 16

Two representations of the DGM and the polygon and ellipse we used to define Target 16 appear below as Figure 74 and Figure 75.

Figure 74 was produced on the data from Oasis Montaj using linear scaling, Channel 1, a minimum millivolt setting of 0 and a maximum millivolt setting of 10. It shows our defined polygon (the geometric basis for our ellipses).

Figure 74. Gridded DGM for Target 16

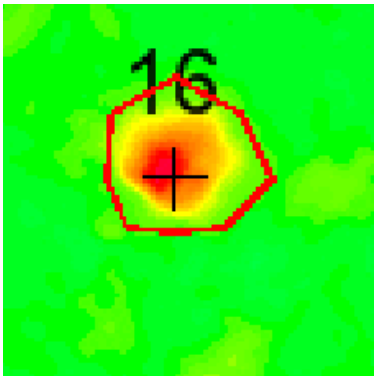
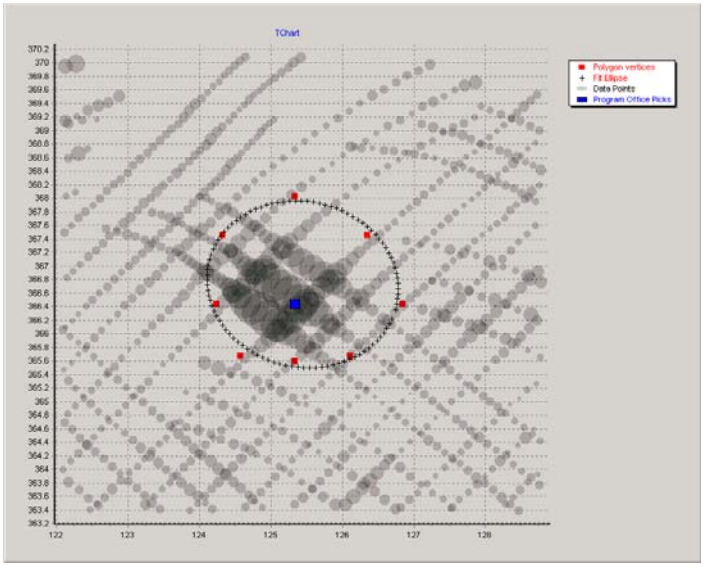


Figure 66 shows the raw data transects for Master ID 16. Each point represents a single DGM reading for channel 1. The size of the point is linearly proportional to the millivolt reading for that point. The minimum millivolt setting was -20 and the maximum millivolt setting was 100. The program office picks are represented by blue squares. The ellipse is our defined target ellipse.

Figure 75. Bubble representation of DGM for Target 16



In retrospect, we could quibble about the data density in the area in the top right of the ellipse. However the data that are there (the SW-NE lines) are regular and we would draw the same ellipse again. It met and exceeded all data density requirements to support the statistics we extracted. So there is nothing apparently odd or unusual about the ellipse definition data density that would explain the false negative.

The only odd thing about Target 16 is the missing field photo to support the ground-truth assertion that it is a 60 mm mortar.

7.2.3 Attribute Space Analysis of False Negatives

This section reviews the position of the false negatives in attribute space, both for iteration one and iteration two attributes. While we will discuss Targets 512 and 444 briefly, they are not particularly interesting as false negatives because (1) Target 512 was correctly classified in our

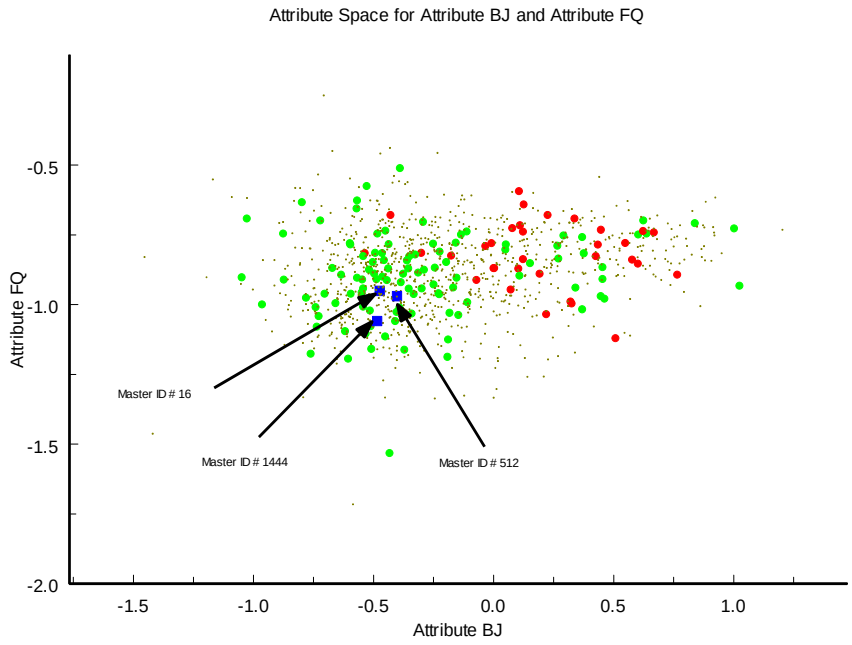
final dig-list iteration (that is the point of iterating); and (2) Target 444 was included in our rankings by mistake—it should have been cannot-analyze.

Thus, the primary emphasis here will be on Targets 16 and 1444, the items that were false negatives in both iterations.

7.2.3.1 Iteration One Attribute Space

All three targets that were false negatives on iteration one were ranked by the LGP ensemble predictor. The attributes used in that predictor were attributes BJ, FQ, and HM. The placement of the three targets within that attribute space are shown in Figure 76 through Figure 78.

Figure 76. Iteration one attribute space for attribute BJ versus attribute FQ (Red circles are UXO. Green circles are Not-UXO. Brown lines are blind data. Iteration one false negatives are highlighted in blue.).



Targets 16, 512 and 1444 are far from the nearest training UXO in attribute space. Had the LGP classifier classified these three targets as UXO on these data, it would have been doing a very poor job of classification. Thus, the issue on iteration one appears to be an unrepresentative sample of targets in the initial training data. We examined the third attribute (HM) and the conclusion is the same.

We next looked at the same attribute space but limited the graphing to just 60mm Mortars. Figure 77 shows that graph.

Figure 77. Attribute space of 60mm mortars for attribute BJ versus attribute FQ (Red circles are false negatives. Blue circles are 60mm mortars in the blind data Green circles are 60mm mortars in the training data)

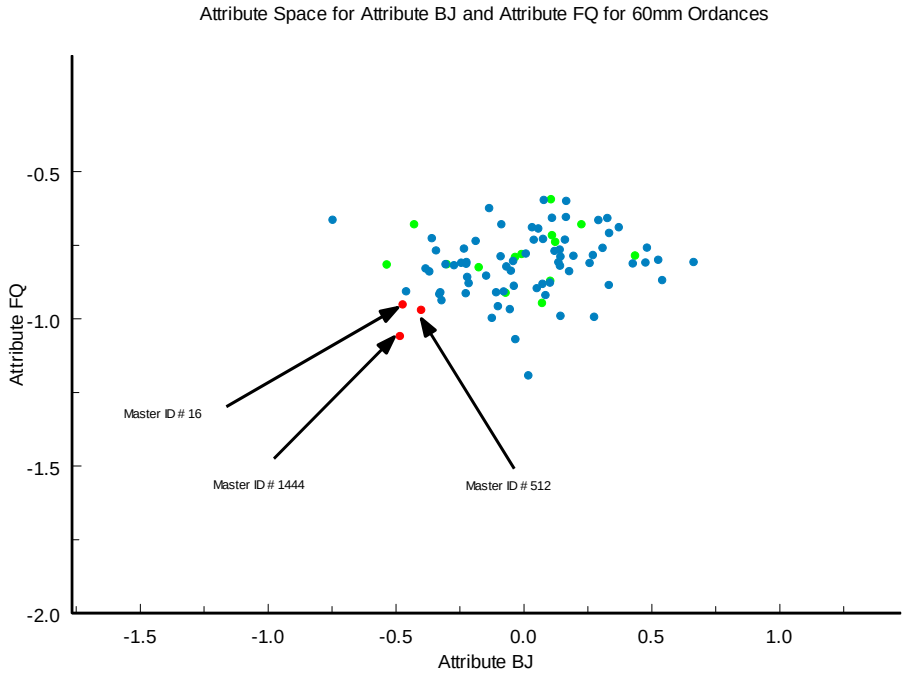
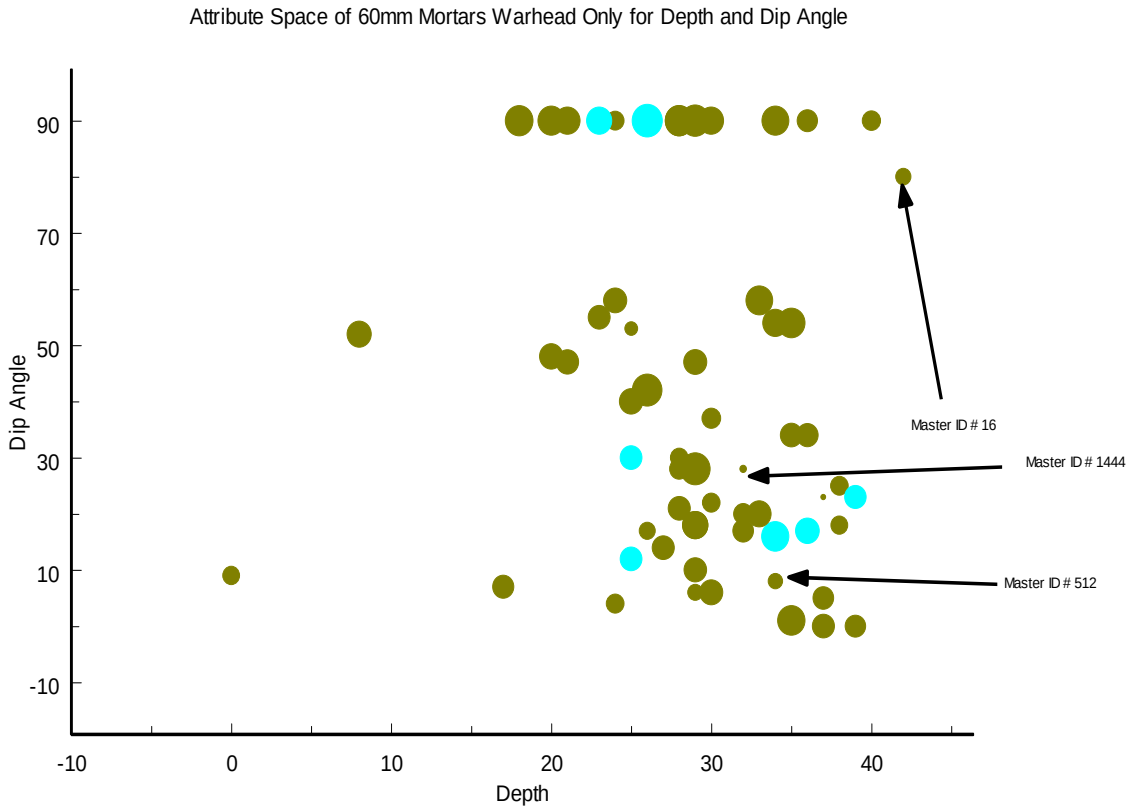


Figure 77 shows the position in attribute space BJ vs. FQ for all 60's on the site. The training data is shown in green, blind in blue. Again, the three false negatives are quite some distance from the nearest training 60mm and are on the outer edge of the cluster that represents all 60's. Also, note that Target 1444 is furthest out from the cluster, which would explain it's low ranking, even relative to Target 16.

Finally, we considered the possibility that the position in attribute space of these three targets could be explained by the physical properties of the targets such as depth and inclination. To test this, we limited the examination to the type of ordnance that comprised false negatives—60mm mortars without tail booms ("small 60's).

Figure 78, plots small 60's on the site by depth and dip angle. So each point represents the position of a single small 60 target by depth and dip angle. The size of each plotted 60mm target was bubbled—that is, the size of the point is proportional to the value of Attribute FQ. Targets that were iteration one training targets are colored blue. Blind targets are colored brown.

Figure 78. All small 60 mm mortars without tail booms. Blue is training data. Brown is blind data. Bubble size gets larger as attribute FQ gets larger.



Three points jump out from Figure 78.

1. Target 512 is significantly smaller on attribute FQ than most other small 60's and is considerably smaller than the nearby targets.
2. Target 1444 has an extremely small value for FQ, regardless of depth and dip angle; and
3. Target 16 is not in a portion of attribute space represented in the iteration one training data. It the deepest small 60mm and almost vertically inclined. The nearest training point is 15 cm shallower than Target 16. However, its value on FQ does not appear to be extreme.

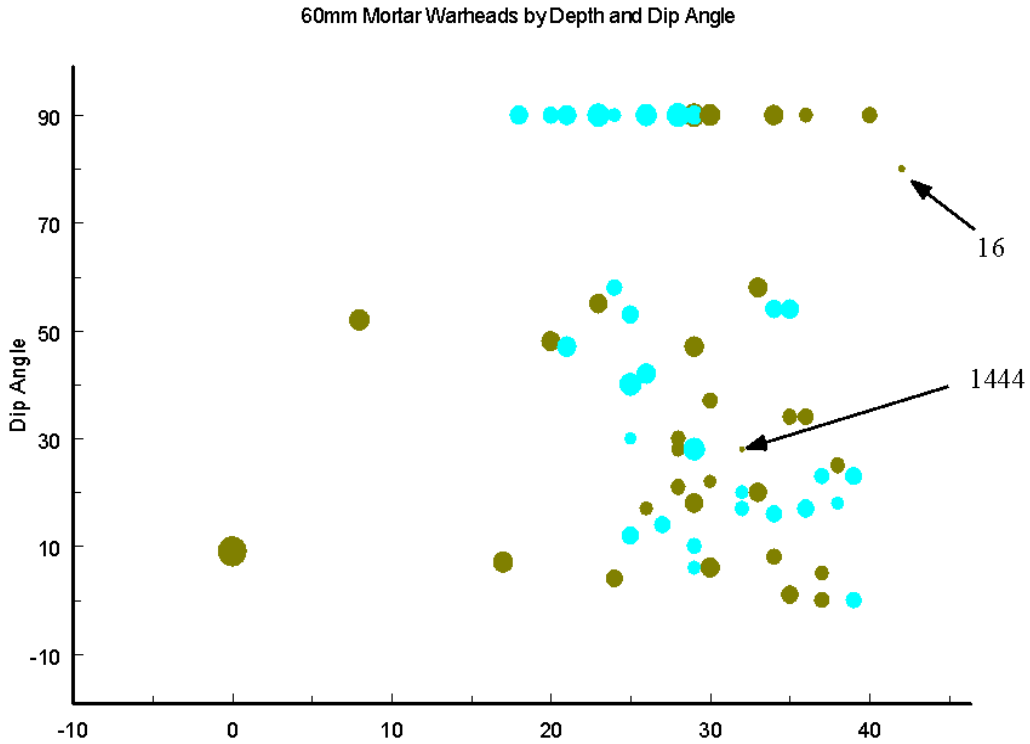
7.2.3.2 Iteration Two Attribute Space

The three false negatives for iteration two were all excluded as high-probability Not-UXO by the amplitude discriminator. Accordingly, this section examines all three targets in the attribute space of the amplitude discriminator—that is Attribute AD2.

The three false negatives on this iteration were Targets 16, 444, and 1444. We have already addressed Target 444 (mistakenly not placed in cannot-analyze). So the focus is on Targets 16 and 1444.

Figure 79 illustrates the same issues highlighted for Targets 16 and 1444 that were addressed regarding the iteration one results. This figure shows only small 60's by depth and dip angle. The size of the bubble for each target gets larger as AD2 (the iteration two amplitude discriminator) gets larger.

Figure 79. Small 60mm mortars by depth and dip-angle. Size of point shows value of target on AD2. Blue is training data. Brown is blind data.



The conclusions to draw from Figure 79 are straightforward: Target 1444 is an extreme outlier on AD2 relative to other small 60's and Target 16 is a somewhat less extreme outlier. Their position on AD2 cannot be accounted for by either depth or dip-angle. We would speculate, therefore, that these two targets have issues with undetected sensor variation.

7.2.4 Correction of Risk-Analysis Procedure Based on Retrospective Analysis

7.2.4.1 Introduction

The retrospective analysis above leads us to ask two questions about our risk analysis procedure as applied in this project:

1. In the experimental plan, we proposed a confidence level for risk analysis of 95%. Would a 99% confidence level be more appropriate?
2. Did we apply our risk analysis procedure correctly when two discriminators are used? We applied risk analysis separately to the amplitude discriminator and the LGP ensemble predictor, using the Bonferonni correction to account for the fact that we were making two confidence assessments. However, we did not note, when doing this, that when we combined the two sets of probabilities, the tail of the risk analysis rankings increased in length. As the stop-digging threshold is set using the cumulative probability of the tail, combining two tails changes the cumulative probability.

Accordingly, we reran the risk analysis at the 95% and 99% confidence levels on the unified dig list, combining the amplitude discriminator and LGP ensemble predictor probabilities before

performing our final risk analysis. It was performed only for the iteration two data as that was the final dig-list. The data was prepared for a unified dig-list as follows:

- First, obtain the Probability of UXO (from the AD2 Risk Analysis already performed) for the targets that fell below the threshold for the Amplitude Discriminator (both for the training and blind data).
- Second, obtain the Probability of UXO (from the LGP Risk Analysis) for the targets that fell above the threshold for the Amplitude Discriminator (both for the training and blind data).
- Third: combine the targets (for the blind and training data) that fell below the amplitude discriminator threshold with the targets that fell above the amplitude discriminator threshold.

At this point, every target has a probability of UXO associated with it.

We then re-ran the kernel regression risk analysis using the combined probability of UXO to set the rank, to obtain a final 'Site' risk analysis.

To determine the site stop-digging threshold the AD2 and LGP individual target probabilities were first converted into ranks across the entire training and blind data. Lower probabilities were interpreted as larger rankings. Next, kernel regression with a Gaussian kernel was used to determine the Probability of UXO for each target:

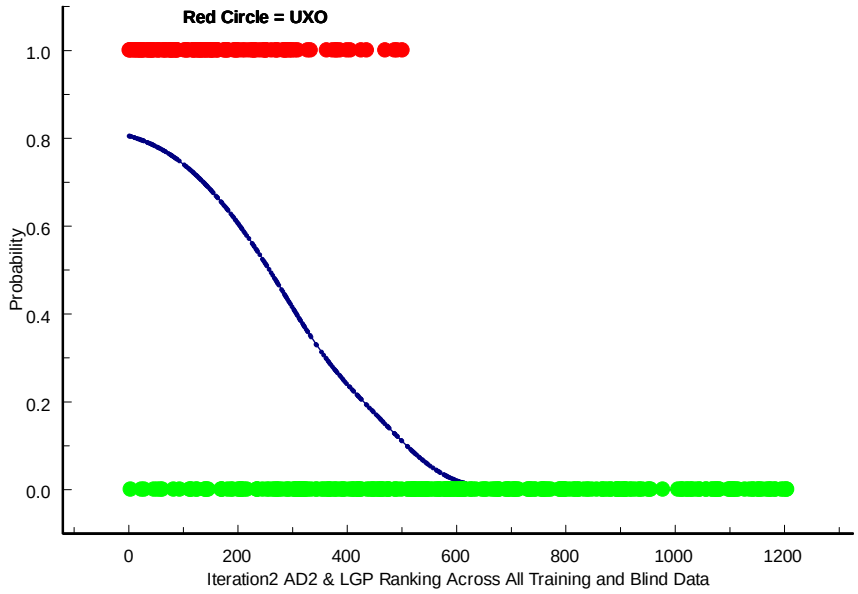
$$\mathrm{P}(UXO)_i = \sum_j e^{-\left(\frac{(x_i - x_j)^2}{2\alpha^2}\right)}$$

Where: (1) α represents the standard deviation of the Gaussian kernel; (2) x_i represents rank of the ith ranked blind data instance computed from the AD2 and LGP individual target probabilities across all training and blind data points; and (3) x_j represents rank of the jth ranked training data instance value of the AD2 and LGP individual target probabilities across all training and blind data points.

The value determined for the parameter, α, is 67.003. That value was determined by n-fold cross-validation on the training data. The α parameter selected was one that produced the minimal value for -2*log likelihood over the training data, which is the maximum likelihood estimator for these data, assuming Bernoulli errors.

Figure 80 shows the derived model plotted against the rankings of the UXO and Not-UXO on the training data. Note that the rankings are derived from the AD2 and LGP individual target probabilities and represent the rankings across all training and blind data not assigned to cannot-analyze one through four.

Figure 80. Retrospective kernel regression fit between UXO and combined AD2/LGP on training data



The Gaussian kernel, generated by the training data, using the above kernel width parameter, was then applied to the ranked blind data, generating a probability that each blind data item is UXO.

Once individual target probabilities are set, the probability that all blind targets above each ranking contain one or more UXO is calculated using the approach outlined in 2.1.6. This is the residual risk as a function of rank. In particular, we used Equation 1 and Equation 2 to compute the OR of the probabilities for all targets from the ranking for which the computation is being performed to the most extremely ranked blind target.

Figure 81 shows the result of applying the kernel model, derived above to the "blind" data (these targets are no longer blind; but we treated them as such in building the kernel model). The blue line in Figure 81 is the probability of UXO as a function of the rank derived from the AD2 and LGP individual target probabilities. The red line is the probability that one or more UXO remain on site at each rank value. When the red line reaches a critical probability value ($p\text{-value}_{crit}$), we assess all targets remaining to the right of that rank (i.e. targets with a larger rank) as high-probability Not-UXO.

Figure 81. Retrospective kernel regression applied to "blind" data

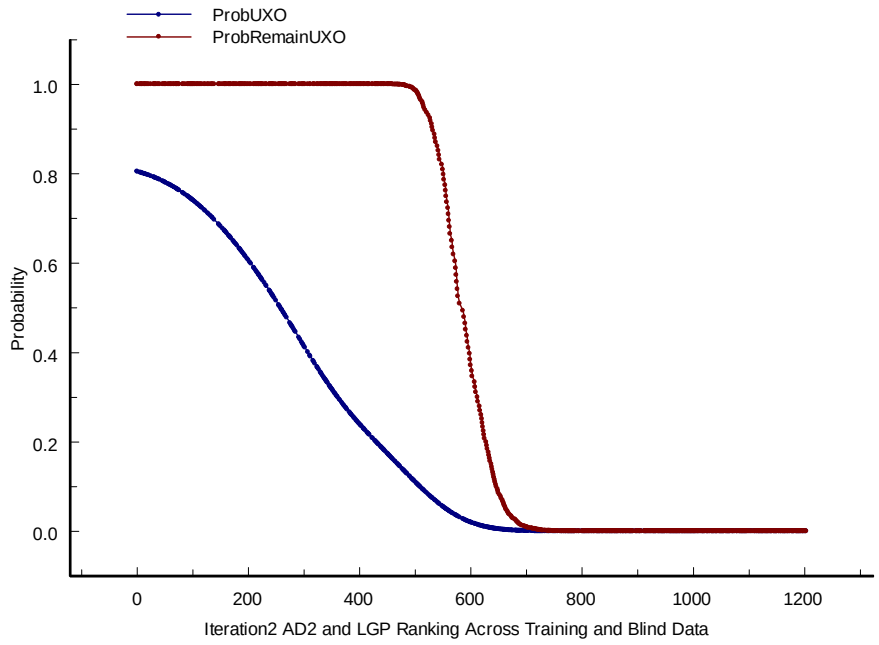


Table 43 shows the results of this risk analysis.

Table 43. Stop-digging thresholds at various confidence levels for retrospective combined risk analysis

Iteration2 AD2 & LGP							
Type	**TID**	**Score**	**HardRank**	**ProbUXO**	**ProbRemainUXO**	**% Of Blind Data Left in Ground**	**# of False Negatives**
Cut-Off Point using 99.99%	**78**	**8.84E-07**	**787**	**5.27E-06**	**4.69E-05**	**25.27%**	**0**
False Negative	1444	8.24E-06	744	6.98E-05	9.20E-04	N/A	N/A
Cut-Off Point using 99%	**958**	**3.12E-05**	**716**	**3.03E-04**	**4.87E-03**	**29.64%**	**1**
False Negative	16	1.39E-04	681	1.47E-03	0.0248	N/A	N/A
Cut-Off Point using 95%	**16**	**1.39E-04**	**681**	**1.47E-03**	**0.0248**	**31.51%**	**2**

At the 95% confidence level, Targets 16 and 1444 remain false negatives. Target 16 is the first target below the stop digging threshold.

At the 99% confidence level, only Target 1444 remains a false negative.

7.2.5 Conclusions Regarding False Negatives

This retrospective analysis yields only suggestions, not definitive answers. For Targets 16 and 1444, there was non-trivial evidence that either that the DGM was not as good as it might have been for those two targets—the measured values for targets 16 and 1444 cannot be explained by any expected variation in the obvious explanatory variables: depth, dip angle, and ordnance type. Even when we control for those variables, Targets 16 and 1444 remain extreme outliers.

We redid our final iteration risk analysis to accommodate the "longer tail" issue highlighted above and to examine different confidence levels. The retrospective risk analysis still yielded two false negatives at the 95% confidence level and one at the 99% confidence level (Target 1444).

It is possible these outliers were caused by an undetected error in the attribute selection or classification processes. But examination of attribute space suggests that the attribute selection process did its job well, given the training data. That is, it selected attributes highly predictive of

98.6% of the UXO (ROC area under the curve of 0.936), including all small and deep 60's—except these two outliers.

Finally, the request for more groundtruth methodology was not at fault. The only way that process would have identified Targets 16 and 1444 would be to have sampled them randomly in the random selection portion of the request because they were the only two representatives of UXO in that portion of attribute space. And the only way to assure that would have been to sample all or a high percentage of targets. But the whole point of discrimination is to minimize random search for UXO. Again, the issue is Targets 16 and 1444 were so unlike other small 60's.

7.3 OBJECTIVE: MAXIMIZE CORRECT CLASSIFICATION OF NON-MUNITIONS

The target was to maximize the number of Not-UXO that were ranked below the final UXO. The objective was 30%. We ranked 28.4% below the final UXO on our final dig-list.

The final ranked UXO was Target 1444. This defined how many Not-UXO could be ranked below it. Target 1444 is an extreme outlier in attribute space and legitimate questions have been raised in Section 7.2 about that Target.

7.4 OBJECTIVE: SPECIFICATION OF NO-DIG THRESHOLD

At our stop digging threshold, 98.6% of the UXO were correctly identified and 35.8% of the Not-UXO remained in the ground. The objectives were 100% and 30% respectively. The 98.6% number (all but three UXO) is discussed at length in Section 7.2.

7.5 OBJECTIVE: MINIMIZE NUMBER OF ANOMALIES THAT CANNOT BE ANALYZED

We determined that the data was sufficient for 82% of the blind targets. The objective was 90%.

There were two principal causes of the difference:

1. We had not seen the data before suggesting this metric. The number of targets in "blobs" was much larger than we had anticipated; and
2. The procedure used to designate target locations where an anomalous region might be a two-peaked target or two nearby targets caused a large number of targets as cannot-analyze category one. See Section 6.5.1.4. The problem was the ambiguity of how the target location was marked. Some were marked using inversion, if the inversion produced a strong and credible fit to that target or targets. Others were selected manually. The demonstrators did not receive any indication which was which. We had to set strict rules for cannot-analyze for these targets in order to assure that training and blind data sets were consistent in the application of that rule. This issue highlighted an area in which our process for designating targets of this type may be improved.

7.6 OBJECTIVE: MINIMIZE THE NUMBER OF BLIND TARGETS SAMPLED

We sampled 20% of the blind targets. The objective was to sample fewer than 20%.

8 FURTHER DISCUSSION OF RESULTS

Three additional topics merit treatment and they are addressed in this section.

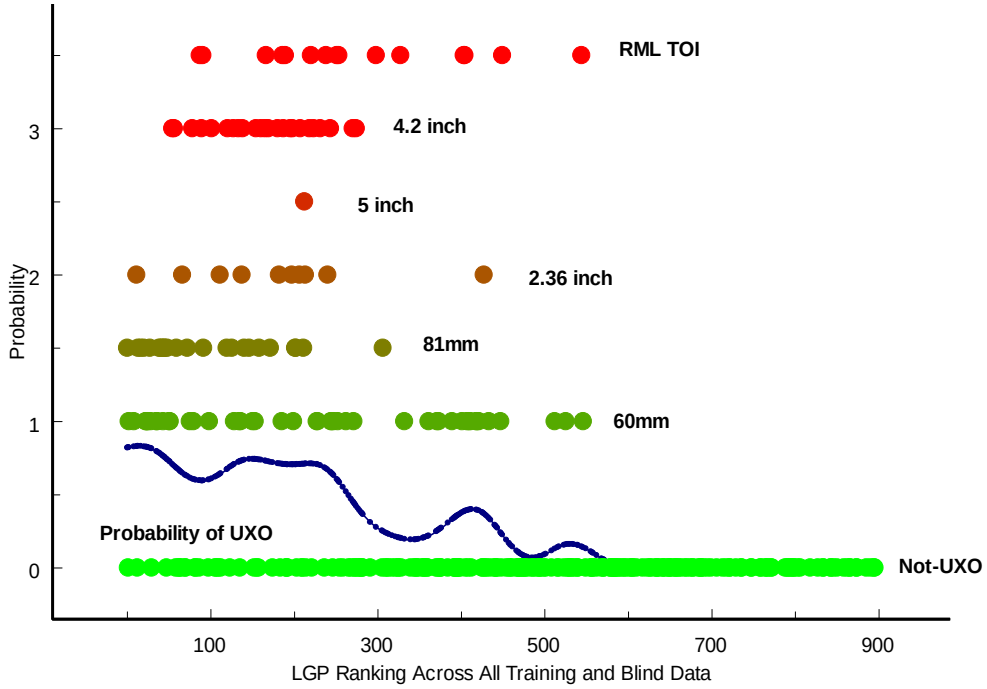
8.1 UNSUPERVISED LGP CLASSIFICATION BY MUNITION TYPE

In iteration two, our LGP ensemble predictor risk analysis produced a sharply and multiply peaked probability curve as a function of rank. Figure 60 and Figure 61 show these peaks.

We investigated the distribution of UXO under those peaks. It was quite clear that LGP was performing non-trivial ordnance type discrimination. Different types of ordnance tended to concentrate under different peaks.

Figure 82 shows the distribution of different ordnance types by LGP ensemble predictor ranks.

Figure 82. Grouping of different UXO types by LGP ensemble predictor



The two peaks on the right were obviously comprised mostly of two groups of 60mm mortars. The wide double peak in the middle was comprised mostly of 4.2 inch mortars and

8.2 EFFECT OF ITERATIVE SAMPLING

Between iterations one and two we sampled additional ground-truth—256 new targets in total. See Section 6.10. In iteration two, our modeling was performed on the original training data plus the new 256 targets sampled. The respective ROC charts for these two iterations are Figure 62 and Figure 63.

In every respect, the second iteration using the larger training set was superior to or equal to the first iteration. Table 44 shows that comparison.

Table 44. Comparison of iteration one and iteration two results

CRITERION	ITERATION ONE	ITERATION TWO
Area Under the Curve	0.858	0.936
Count of Not-UXO Left in Ground after Last UXO	124	364
Percent Not-UXO Left in Ground after Stop-Digging	27.59%	35.88%
False Negatives	3	3
False Negatives other than Mistaken Cannot Analyze	3	2

The increase in area under the curve from iteration one to iteration two is very substantial. The error implied by the area under the curve (1-AUC) is more than halved.

In addition, the count of Not-UXO ranked below the final UXO approximately triples while the amount of Not-UXO ranked lower than the stop digging threshold increases by about 30%.

In short, the acquisition of new groundtruth in the Request for more Groundtruth improved the UXO classification significantly.